ELEMENTS DE LINGUISTIQUE ET DE PRAGMATIQUE POUR LE TRAITEMENT AUTOMATIQUE DU LANGAGE :

Du signe au sens

Jean Caelen
DR-CNRS

Table des matières

Introduction 3
La sémiotique 4
Le signifiant et le signifié : signe dyadique (Saussure, Hjelmslev) 4
Le carré sémiotique et le schéma actantiel (Greimas) 5
Le signe ternaire de Pierce 6
Le lexique 8
Les catégories lexicales 8
Les dimensions du lexique 10
La morphosyntaxe 11
Les bases de données lexicales pour le TALN 13
Les ontologies linguistiques 15
La syntaxe 17
La grammaire générative de Chomsky 17
Les grammaires lexicales fonctionnelles de Bresnan-Kaplan 17
Les grammaires catégorielles (GPSG) de Gazdar et grammaires dérivés (HPSG, TAG) 20
Les grammaires cognitives 25
Les modèles de langage (LLM) probabilistes 26
La sémantique 34
La sémantique descriptive 34
La sémantique générative 35
La sémantique interprétative : grammaire de cas de Fillmore 37
La sémantique logique 38
La sémantique distributionnelle 38
La sémantique cognitive 40
Le web sémantique 43
La rhétorique 45
La pragmatique 49
La référence 51
Les actes de langage 54
Applications 59
Résumé de texte 59
Générateur de texte 60
Traduction automatique 61

Introduction

Le langage est au cœur de l'expérience humaine : il permet de communiquer, de raisonner, de transmettre des connaissances et de construire des cultures. Depuis les premiers travaux en intelligence artificielle menés par des chercheurs comme Alan Turing ou Noam Chomsky, une question fondamentale se pose : **comment faire comprendre le langage humain aux machines ?**

Le **traitement automatique du langage naturel (TALN)**, souvent appelé **NLP** (Natural Language Processing), est un domaine interdisciplinaire à la croisée de la linguistique, de l'informatique et des statistiques. Il vise à concevoir des méthodes permettant aux ordinateurs d'analyser, de comprendre, de générer et d'interagir en langage humain. Aujourd'hui, le TALN est omniprésent : moteurs de recherche comme Google, assistants vocaux comme Siri, traduction automatique comme DeepL ou systèmes conversationnels tels que ChatGPT reposent tous sur des techniques de traitement du langage.

Au fil des décennies, le TALN est passé de méthodes fondées sur des règles linguistiques à des approches statistiques, puis à l'apprentissage profond et aux modèles de grande taille (LLM). Cette évolution a profondément transformé notre capacité à traiter des textes et à construire des systèmes capables de dialoguer, résumer des documents, détecter des sentiments ou répondre à des questions.

Dans ce cours, nous explorerons les concepts fondamentaux du traitement automatique du langage :

- La représentation des mots et des phrases,
- Les modèles logiques, graphes, probabilistes et neuronaux,
- L'analyse syntaxique et sémantique,
- Les notions principales en sémiotique, rhétorique et pragmatique
- Les applications concrètes du TALN.

L'objectif de ce cours est double : **comprendre les principes théoriques du TAL** et **acquérir les compétences pratiques nécessaires pour concevoir et évaluer des systèmes linguistiques automatisés**. En étudiant le langage à travers les machines, nous apprendrons aussi à mieux comprendre la structure et la richesse du langage humain lui-même.

La sémiotique

Un signe est la manifestation phénoménologique d'une chose ou d'un phénomène qu'il exprime de manière plus ou moins explicite. De façon générale, on peut dire qu'un signe est un objet porteur d'une signification autre que sur lui-même (mais en même temps sur lui-même sur le plan de sa constitution propre à savoir qu'un dessin de cheval est aussi un dessin). Le signifié ne peut être séparé du signifiant. La sémiotique n'est pas seulement une théorie du signe mais une théorie de la signification : la signification est « le produit organisé par l'analyse » ; elle est toujours articulée.

La sémiotique est l'étude des relations entre un signe et un référent (objet, concept, phénomène, etc.). La relation peut être considérée comme une relation en-soi (relation signifié/signifiant) ou comme une relation pour-soi (la relation signifiant/référent pour un interprétant) ce qui a donné deux écoles de pensée : la première issue de De Saussure puis de l'école de Paris est dite dyadique et la seconde issue de Pierce est dite triadique. Ces deux écoles se rattachent pour l'une au structuralisme et pour l'autre à la logique et plus largement au pragmatisme logique.

On peut considérer le signe selon les deux points de vue (a) structural (dyadique) ou (b) fonctionnel (triadique) :

(a) Du côté de la structure on distingue le plan de l'expression (le signifiant) et celui du contenu (le signifié), la forme, ce qui structure, et la substance, ce qui est structuré. C'est le signe-relation.

(b) Du côté de la fonction cette relation se projette sur un interprétant (qui n'est pas un interprète mais un processus d'interprétation). C'est le signe-action.

Charles Morris distingue lui, trois "dimensions" de la sémiotique :

1. Sémantique : la relation entre les signes et ce qu'ils signifient (relations internes entre signifiant et signifié ou relation externe entre le signe global et le référent).
2. Syntaxe : les relations entre signes.
3. Pragmatique : la relation entre les signes et leurs utilisateurs.

Morris voulait développer une science des signes « sur une base biologique et particulièrement dans le cadre de la science du comportement ». Le linguiste et sémioticien Thomas Sebeok fut l'un de ses étudiants et inaugura la voie de la biosémiotique reprise de nos jours par une branche de la biologie (écoles de Tartu et de Copenhague).

Le signifiant et le signifié : signe dyadique (Saussure, Hjelmslev)

Pour De Saussure le signe est double :

Le **signifiant** (le mot porteur de sens comme "chien", "dog"),

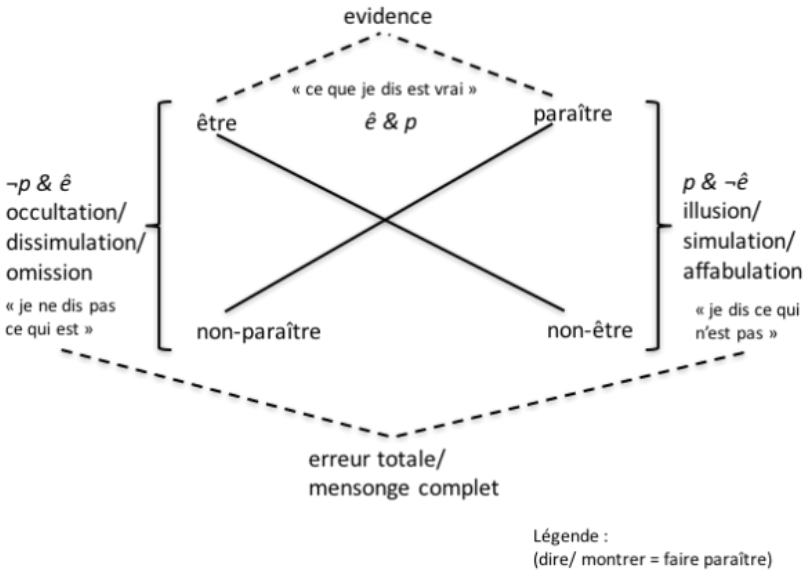
Le **signifié** (la chose nommée, "chien" référant à un chien particulier).

Le signe linguistique est formé d'un signifiant et d'un signifié : il est dyadique et **arbitraire**.

Selon Hjelmslev, la langue est un système de signes. Le signe, que Hjelmslev appelle

glossème, possède, selon lui, un contenu et une expression. Chacun de ces deux termes a une forme et une substance :
Glossème -> Contenu + Expression
Contenu -> Substance (le référent qui appartient à la réalité)
+ Forme (plérème ou signifié selon Saussure)
Expression -> Substance (le son)
+ Forme (cénème, phonème ou signifiant selon Saussure)

Le carré sémiotique et le schéma actantiel (Greimas)



Le carré sémiotique est un carré logique qui permet de mettre les signes en relation par un système d'oppositions binaires. Dans l'exemple ci-contre l'être (S1) s'oppose au paraître (S2) sur une dimension sémiotique et l'être s'oppose au non-être (¬S1) sur une dimension logique, de même pour le paraître et le non-paraître (¬S2).

S'appuyant sur ce carré sémiotique, Greimas propose une sémiotique générale qui privilégie diverses manifestations du récit : mythes, légendes, contes, nouvelles, romans, recettes de cuisine, etc. Elle s'est surtout attardée, dans les années 1960-1970, au parcours génératif de l'action (du provenir au parvenir) ; mais depuis, elle s'est davantage intéressée à la passion (de l'advenir au survenir) et à la cognition, la dimension thymique (tensive, passionnelle) et la dimension cognitive prenant le dessus sur la dimension pragmatique et les catégories sur les dimensions. La mise en mouvement du signe s'opère autour du schéma actantiel qui vise à décrire les rôles et les relations entre les personnages (ou actants) dans un récit, indépendamment de leur nature concrète (humains, animaux, objets, idées, etc.).

Les six actants principaux
Greimas propose un modèle basé sur **six rôles actantiels** fondamentaux, organisés en trois paires d'opposition :

Actant	Rôle
Sujet	Celui qui cherche à accomplir une quête ou un objectif.
Objet	Ce que le sujet désire obtenir ou atteindre.
Destinateur	Celui qui envoie le sujet en quête (peut être une personne, une règle, une valeur).

Actant	Rôle
Destinataire	Celui qui bénéficie de la quête (peut être le sujet lui-même ou un tiers).
Adjuvant	Celui qui aide le sujet dans sa quête.
Opposant	Celui qui s'oppose au sujet et à la réalisation de sa quête.

Exemple d'application
Prenons l'exemple du conte **"Le Petit Chaperon rouge"** :

- **Sujet** : Le Petit Chaperon rouge
- **Objet** : Porter une galette et un pot de beurre à sa grand-mère
- **Destinateur** : La mère (qui envoie le Petit Chaperon rouge)
- **Destinataire** : La grand-mère (et indirectement, le loup)
- **Adjuvant** : Le chasseur (qui sauve la grand-mère et le Petit Chaperon rouge)
- **Opposant** : Le loup

Intérêt du schéma actantiel

- **Universalité** : Il permet d'analyser des récits très variés (contes, romans, films, publicités, etc.).
- **Abstraction** : Il ne se limite pas aux personnages humains, mais peut s'appliquer à des forces, des idées, des objets.
- **Structure** : Il met en lumière la dynamique narrative et les enjeux du récit.

Limites et critiques

- **Réductionnisme** : Certains récits complexes ne se laissent pas facilement réduire à six actants.
- **Variantes** : D'autres théoriciens ont proposé des modèles complémentaires ou alternatifs (comme Vladimir Propp pour les contes).

Le signe ternaire de Pierce

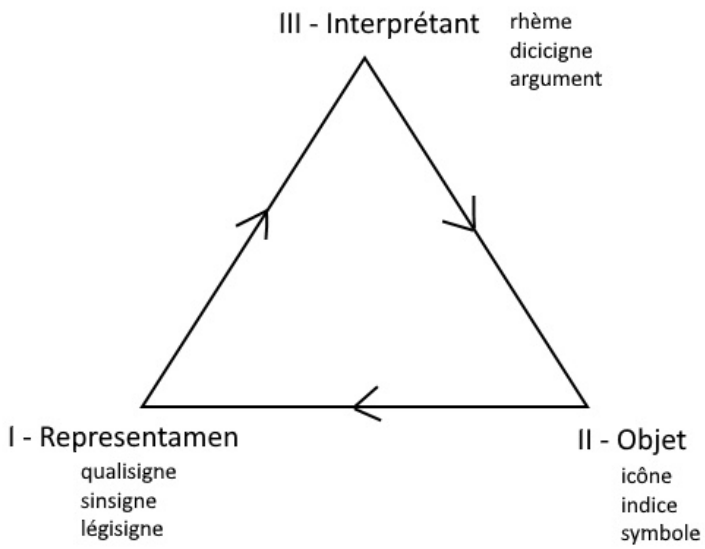
Le processus de sémiose (sémiosis) présente un ordre logique :

- Un representamen (signe concret) est premier (**primeité**) ;
- Un interprétant (esprit qui interprète le signe) est troisième (**tiercéité**) ;
- Un objet (dynamique) est second (**secondéité**). L'objet immédiat est incorporé dans le representamen.

La conception triadique peircienne du signe est représentée sous forme d'un triangle dont chaque pointe se décline à nouveau en priméité, secondéité, tiercéité.

Les trois dimensions :

(a) syntaxique : le signe en lui-même,
(b) sémantique (existentielle) : le signe dans son rapport à l'objet,
(c) pragmatique : le signe dans son rapport à l'interprétant.



Dimensions	Catégories		
	priméité	secondéité	tiercéité
/lui-même	*qualisigne*	*sinsigne*	*légisigne*
/objet	*icône*	*indice*	*symbole*
/interprétant	*rhème*	*dicisigne*	*argument*

Ex : l'icône ressemble à son objet (le panneau routier), l'indice est un existant actualisé (le cadran de l'horloge), le symbole exige des inférences intellectuelles pour conduire à l'objet ("oiseau").

La sémiose peut être vue comme un processus qui, à partir d'un *representamen*, conduit un continuum d'interprétants à produire un objet. Les interprétants sont à chaque étape des representamens : l'interprétant du representamen devient à son tour representamen du même objet pour un nouvel interprétant, et ceci indéfiniment. Dit autrement, le signe representamen est traduit (par inférence) par l'interprétant dans un nouveau signe plus complexe qui est une partie du sens. On voit 3 en premier mais en fait il est dernier (c'est le representamen qui est premier). Par endroit, Peirce indique que 2 détermine 3 à travers 1.

Pour Peirce, toute pensée est manipulation de signes et aucune pensée ni signe n'existe indépendamment d'autres pensées ou signes. Si on considère la relation du signe à son objet (pointe II du triangle) : un objet s'impose à un sujet via une icône (priméité). Le sujet réagit à cette icône par une rupture et une réaction sous forme d'indice (secondéité). Pour atténuer les affects, comprendre la situation, prendre du recul, un interprétant ou symbole (tiercéité) va élaborer une (des) loi(s) concernant la dyade icône/indice.

Le lexique

Les catégories lexicales

On distingue deux grandes classes lexicales :

Mots outils (ou mots grammaticaux)	**Mots lexicaux**
Déterminants	Adverbes
Pronoms	Noms
Prépositions	Adjectifs
Conjonctions	Verbes
Interjections	Locutions particulières (lexicalisées)

La première catégorie est peu évolutive et constante dans presque tous les champs d'application. La seconde est très variable en ce qui concerne la quantité des instances des classes : c'est la richesse du vocabulaire.

- **Les déterminants**

  Les articles définis (nom dont le sens n'est pas déterminé) : *le, la, l', les*.
  Les articles indéfinis (objet nommé non précisé ou non connu) : *un, une, des*.
  Les articles contractés (fusion des prépositions *à,* et, *de,* avec, *le* et *les*) : *au (à + le), aux (à + les), du (de + le), des (de + les)*.
  Les articles partitifs (partie/quantité indéfinie de l'objet) : *du, de la, de l', (des)*.
  Les déterminants possessifs singulier : (à moi, à toi, à lui, à elle, à soi) : *mon, ton, son ; ma, ta, sa ; mes, tes, ses*.
  Les déterminants possessifs pluriel : (à nous, à vous, à eux, à elles) : *notre, votre, leur ; nos, vos, leurs*.
  Les déterminants démonstratifs : ce, cet, cette, ces
  Les déterminants exclamatifs/interrogatifs : *quel, quelle, quels, quelles*.
  Les déterminants indéfinis : *aucun(e)(s), chaque, plusieurs, quelque(s), certain(e)(s), etc.* et locutions : *beaucoup de, énormément de, peu de, etc.*
  Les déterminants numéraux : *trois, troisième, cent, mille, premier, zéro, etc.*

- **Les pronoms**

  Les pronoms personnels : *je, tu, il, elle, on, nous, vous, ils, elles ; le, la, les, l', me, te, se, lui, leur.*
  Les pronoms possessifs : *le mien, le tien, le sien ; la mienne, la tienne, la sienne ; les miens, les tiens, les siens ; les miennes, les tiennes, les siennes ; le nôtre, le vôtre, le leur ; la nôtre, la vôtre, la leur ; les nôtres, les vôtres, les leurs.*
  Les pronoms interrogatifs : *laquelle, duquel, qui, que, auquel…*
  Les pronoms démonstratifs : *ce, c', cela, ça, ceci, celui, celle, ceux, celles.*
  Les pronoms indéfinis : *certain(e)s, aucun(e), chacun(e), personne, rien, tout, tous…*

Les pronoms relatifs : *qui, que, dont, où, lequel, laquelle, auquel, duquel, desquelles…*

- **Les prépositions**

à, après, attendu que, au-dedans, avant, avec, chez, concernant, contre-courant, dans, de, dedans, dehors, depuis, derrière, dès, dessous, dessus, devant, dixit, durant, en, entre, envers, environ, excepté, hormis, hors, joignant, jusque, malgré, moyennant, nonobstant, outre, par, par-delà, par-dessus, parce que, parmi, pendant, pour, sans, sauf, selon, sous, sur, tandis, vers, versus, via, vu, etc.

- **Les conjonctions**

Les conjonctions de coordination : *mais, où, et, donc, or, ni, car*
Les conjonctions de subordination :
Simple : *que, si, comme, lorsque, quand, quoique, puisque.*
Le temps : *quand, dès que, pendant que, avant que, après que, lorsque, alors que…*
Le but : *afin que, pour que, de sorte que…*
La cause : *comme, parce que, sous prétexte que…*
La comparaison : *ainsi que, comme, de même que, autant que…*
La condition : *à condition que, pourvu que, si…*
La conséquence : *au point que, de manière que, si bien que…*
L'opposition : *alors que, quand, pendant que, tandis que…*
La concession : *même si, bien que, quoique…*

- **Les interjections**

Psst ! heu ! ah ! oh ! hum ! ah ! ouf ! eh ! tiens ! oh ! peuh ! pouah ! ouïe ! ouh ! aïe ! eh ! hein ! heu ! pfuitt ! Tiens ! eh ! peuh ! oh ! heu ! bon !

- **Les adverbes**

Manière *Adj+[ment]*
Lieu *ici, devant, derrière*
Temps *hier, demain*
Quantité
  Modificateurs
    Comparatifs *plus, moins, autant, aussi*
    Superlatifs *très, trop, peu, beaucoup, très bien,*
  Interrogatifs, exclamatifs *comment, où, quand*
Négation *pas, point, guère, plus, jamais, personne, rien, aucun, nul*
Opinion *oui, si, non*
Modalisateurs
  Insistance *certainement*
  Réserve *probablement, peut-être*
Liaison *ensuite, puis, ainsi, en effet, aussi*

Locutions adverbiales de coordination *à l'improviste*
Interrogatifs
Lieu *où*
Cause *pourquoi*
Temps *quand*
Manière *comment*
Quantité *combien*

- **Les noms ou substantifs**

Peuvent être classés dans une matrice de traits : commun/propre, animé/non-animé, humain/non-humain, concret/abstrait, comptable/non-comptable, simple/composé, individuel/collectif, etc.

- **Les verbes**

Le temps, le mode, la personne, la voie (actif, passif, réflexif), l'aspect

Le procès (performatif)
Factitif (peut être remplacé par 'faire'+infinitif)
mouvement : *avancer,*
manipulation : *prendre, écrire*
Perfectif/imperfectif (passé encore actuel ou non)
comprendre/posséder
Duratif/momentané *itérer/copier*

L'aspect donne une précision de sens au verbe :
Accompli/non accompli (déjà fait/à faire)
Inchoatif (débute ou finit)
Progressif (être en train de)
Immédiat (être sur le point de)
Résultatif

- **Les adjectifs et participes**

Qualifiant : taille, couleur, sentiment, etc.
Classifiant : catégorie, type, appartenance, etc.
Adjectifs verbaux (participes présent ou passé)

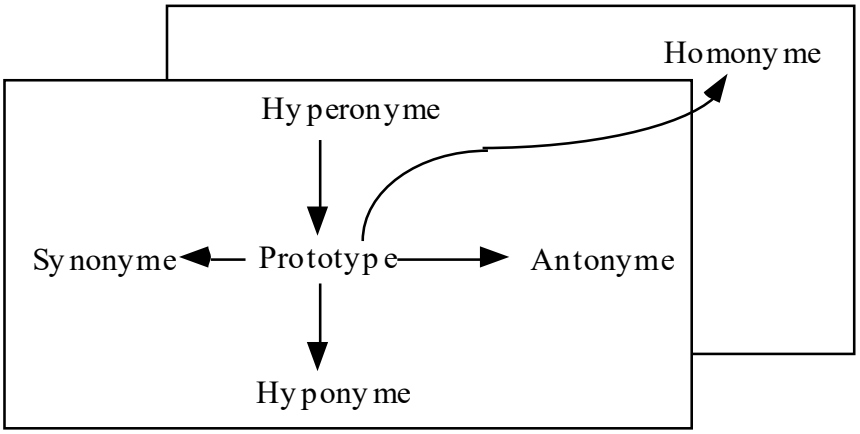
Les dimensions du lexique

La sémantique lexicale
La synonymie est une relation d'équivalence permettant de substituer un terme (ou un segment) à un autre terme (ou segment), sans modifier le sens global de l'énoncé. En tant que relation lexicale, la synonymie n'a malheureusement pas les bonnes propriétés des relations mathématiques d'équivalence, simplement parce qu'à cause de la polysémie,

la propriété de transitivité n'est pas vérifiée.

Habituellement, l'antonymie est définie comme une notion d'incompatibilité entre deux termes. À la lumière de la représentation vectorielle (voir ci-après), nous préférons plutôt considérer une notion de symétrie qui se définit comme suit : deux termes sont en relation d'antonymie si on peut exhiber une symétrie de leurs traits sémantiques par rapport à un axe. L La symétrie peut se décliner de différentes manières selon la nature de son support. On distingue, comme supports de symétrie :

- Une propriété affectant une valeur étalonnable (valeur élevée, valeur faible) : par exemple, *chaud, froid* sont des valeurs symétriques de température, sur une échelle implicite.
- L'application d'une propriété (applicable/non applicable, présence/absence) : par exemple, informe est antonyme de tout ce qui a une forme, insipide, incolore, inodore, etc. de tout ce qui pourrait avoir saveur, couleur, odeur,
- L'existence d'une propriété ou d'un élément considéré comme symétrique par l'usage (e.g. *soleil/lune*), ou par des propriétés naturelles ou physiques des objets considérés (e.g. *mâle/femelle, tête/pied* . . .). Les antonymes vont alors par paires



La morphosyntaxe

La **morphosyntaxe** concerne l'ensemble des structures qui permettent de construire grammaticalement un énoncé. Elle porte aussi bien sur les formes des mots, flexions régulières et irrégulières, variantes irrégulières de certains noms et verbes, l'agencement des marques syntaxiques autour du nom (déterminants, etc.), du verbe (pronoms, etc.), de l'adjectif, de l'adverbe, et enfin de l'organisation des mots et groupes de mots dans un énoncé ou une phrase. Dans la langue française, tous les niveaux d'organisation langagière sont touchés de manière importante par la morphosyntaxe. On distinguera quatre niveaux de morphosyntaxe : lexical (racine des mots), flexionnel (terminaison des mots), contextuel (marqueurs syntaxiques ayant un caractère obligatoire et dont l'emplacement est strictement déterminé) et positionnel (organisation des mots ou groupes de mots présentant une certaine flexibilité). Ces quatre niveaux d'organisation correspondent le plus souvent à l'âge des structures langagières et à leur évolution au cours du temps, des plus anciennes (lexicales) au plus récentes (positionnelles). En revanche, l'utilisation est largement indépendante de l'âge des structures et tous les

niveaux interagissent dans la morphosyntaxe du français actuel.

L'utilisation du concept de morphosyntaxe permet de s'affranchir de ce clivage entre lexique et syntaxe. La morphosyntaxe comporte des éléments et des structures qui modifient les éléments lexicaux de manière plus ou moins proche du radical du mot et qui agencent les éléments lexicaux ainsi modifiés pour créer un énoncé complet. Ainsi, si l'on prend l'exemple du verbe être en français, les personnes du verbe peuvent être gérées de trois manières différentes : 1. sur le radical : /syi/ (suis, 1ère personne singulier) vs. /e/ (es, 2ème personne singulier). 2. sur la flexion : /səʀɔ̃/ (serons, 1ère personne pluriel) vs. /səʀe/ (serez, 2ème personne pluriel). 3. sur le pronom personnel1 : /ty e/ (tu es, 2ème personne singulier) vs. /il e/ (il est, 3ème personne du singulier.

Dans l'exemple ci-dessus, le pronom personnel est ce qu'on appelle un *clitique*. Il s'agit d'une forme qui ressemble à un mot, qui est séparée du verbe, entre laquelle on peut insérer des éléments en nombre limité (négation, autres pronoms personnels) et qui a un caractère obligatoire (on ne peut l'omettre sans que la forme devienne agrammaticale, sauf en français dans les cas d'impératif ou d'existence d'un groupe nominal sujet). Le clitique fait traditionnellement partie de la syntaxe tandis que les variations flexionnelles font partie de la morphologie syntaxique tandis que les variations de radical appartiennent à la morphologie lexicale. L'ensemble relève de la morphosyntaxe et peut s'unifier dans un même cadre.

Morphosyntaxe lexicale exemple Variation de temps pour un même verbe (être) suis – serai – fus – étais Variation de catégorie grammaticale dormir – somme

Morphosyntaxe flexionnelle
Changements catégoriels exemple Variation de catégorie modifier – modification grammaticale marteau – marteler
Flexions nominales exemple Variation de genre joueur – joueuse Variation de sens bosse – bosselage – bossellement – bosselure – bossette – bossu
Flexions adjectivales exemple Variation de genre vert – verte joli – jolie
Flexions verbales exemple Variation de temps danse – dansait – dansera Variation de personne danse – dansons – dansez

Morphosyntaxe contextuelle
Le groupe nominal exemple Variation de genre le tour – la tour Variation de nombre l'enfant – les enfants Variation de rôle papa– à papa– pour papa
Le groupe verbal exemple Variation de temps a dansé – avait dansé – aura dansé Variation de personne je danse – tu danse – il danse – elle danse – on danse Variation d'aspect je danse –j'ai dansé – je vais danser Variation modale je danse – je veux danser – je dois danser – je peux danser – je fais danser Négation je danse – je (ne) danse pas

> *Note : Les variations aspectuelles sont souvent considérées comme les variations verbales les fondamentales dans la plupart des langues. Il s'agit de variations qui portent sur la nature du « procès » décrit par le verbe à un moment donné, vu de manière interne. Le procès en linguistique correspond à l'ensemble des valeurs sémantiques d'un verbe (ou prédicat verbal). Par exemple, le procès peut être*

> *statique ou dynamique, et s'il est dynamique il peut être en cours, achevé, sur le point de commencer.*

Morphosyntaxe positionnelle exemple Variation positionnelle Chirac a battu Jospin – Jospin a battu Chirac

Les bases de données lexicales pour le TALN

Les ressources linguistiques sous forme de dictionnaires

Le portail lexical est un projet mené par le laboratoire ATILF dont l'objectif est de valoriser des ressources linguistiques issues de différents projets de recherche au sein d'un portail unique.

Les ressources actuellement utilisées pour le portail lexical sont :

Morphologie	Le lexique Morphalou 2.0 Phonétisation réalisée par l'équipe Parole au LORIA Le synthétiseur de voix utilise LIA_PHON et MBROLA
Lexicographie	Le Trésor de la Langue Française informatisé (TLFi) Les dictionnaires de l'Académie Française (4ème, 8ème et 9ème édition) La Base de données lexicographiques panfrancophone de l'université Laval de Québec La Base Historique du Vocabulaire Français du laboratoire ATILF Le Dictionnaire du Moyen Français (1330 - 1500) du laboratoire ATILF Le Du Cange (Moyen Âge) de l'École Nationale des Chartes
Etymologie	Le Trésor de la Langue Française informatisé Le projet de recherche TLF-Étym (mise à jour des notices étymologiques du TLF)
Synonymie et Antonymie	Le dictionnaire de synonymes du laboratoire CRISCO
Proxémie	projet Prox du laboratoire ERSS
Concordance	Le corpus Frantext libre de droits

Représentation vectorielle

Le modèle vectoriel n'est pas récent, puisqu'il a été au départ introduit par Salton en informatique documentaire [Salton 1968]. Sa réhabilitation dans les recherches en TALN est en revanche relativement récente, car elle a été essentiellement motivée par la mise à disposition des chercheurs de grandes bases de textes grâce au web en particulier, alors que précédemment, ces recherches passaient par des phases ardues de constitution de corpus d'expérience.

Le modèle de vecteurs conceptuels s'appuie sur la projection dans un modèle mathématique de la notion linguistique de champ sémantique. Tout terme (lexie) et tout concept est projetable sur les vecteurs de la famille génératrice, et est donc représenté par un vecteur conceptuel. Mieux encore, on peut calculer le thème de tout segment de

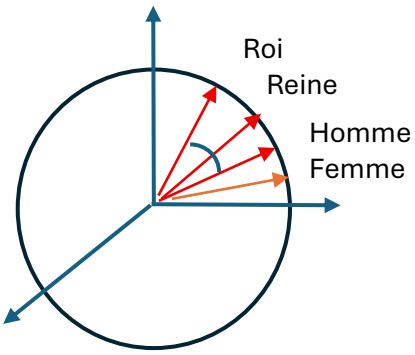
texte tel que documents, paragraphes, syntagmes, etc. sous forme de vecteur conceptuel : c'est le sens du segment en question. Cette représentation homogène du sens, quelle que soit la granularité, est très avantageuse pour la classification des textes, l'indexation et la recherche évoluée d'information.

Chaque mot est représenté par un vecteur normalisé (longueur = 1) de dimension d. Ainsi l'ensemble du lexique peut être projeté sur une hypersphère. La proximité entre deux mots est mesurée par le cosinus de l'angle des deux vecteurs représentatifs. On peut s'arranger pour avoir des champs lexicaux homogènes où les distances entre mots mesurent la proximité sémantique. Par exemple : soit C un ensemble fini de n concepts. Un vecteur conceptuel V est une combinaison linéaire des éléments c_i de C. Pour un sens A, le vecteur V_A est la description (en extension) des activations des concepts de C. Par exemple, les sens de *ranger* et de $_{couper}$ peuvent être projetés sur les concepts suivant (les CONCEPT[intensité] étant ordonnés par intensité décroissante) :

V_{ranger} = (CHANGEMENT [0.84], VARIATION [0.83], ÉVOLUTION [0.82], ORDRE [0.77], SITUATION [0.76], STRUCTURE [0.76], RANG [0.76], etc.).
V_{couper} = (JEU [0.8], LIQUIDE [0.8], CROIX [0.79], PARTIE [0.78], MÉLANGE [0.78] FRACTION [0.75], SUPPLICE [0.75], BLESSURE [0.75], BOISSON [0.74], etc.).

Il est souhaitable de pouvoir mesurer la proximité entre les sens représentés par deux vecteurs (et donc celle de leur mot associé). Soit Sim(X, Y) la mesure de similarité, utilisée habituellement en recherche d'informations, entre deux vecteurs, définie selon la formule (1) ci-dessous (avec "·" étant le produit scalaire). On notera que l'on suppose ici que les composantes des vecteurs sont toujours positives ou nulles (ce qui n'est pas nécessairement le cas). Enfin, nous définissons une fonction de distance angulaire D_A entre deux vecteurs X et Y selon la formule (2).

(1) $Sim(X, Y) = \cos(X, Y) = X \cdot Y$ (X et Y sont normés)
(2) $D_A(X, Y) = Arcos(Sim(X, Y))$



Intuitivement, cette fonction constitue une évaluation de la proximité thématique et est en pratique la mesure de l'angle formé par les deux vecteurs. On considérera, en général, que pour une distance $D_A(X, Y) \leq \pi/4$ (soit 45 degrés), X et Y sont sémantiquement proches et partagent des concepts par exemple *roi* et *homme*, fig. ci-contre). Pour $D_A(X,Y) \geq \pi/4$, la proximité sémantique de A et B sera considérée comme faible. Aux alentours de $\pi/2$, les sens sont sans rapport. La synonymie (dans son acception la plus large) est incluse dans la proximité thématique, cependant elle exige, de plus, la concordance des catégories morphosyntaxiques. L'inverse n'est évidemment pas vrai.

La distance angulaire est une vraie distance (contrairement à la mesure de similarité) et elle vérifie les propriétés de réflexivité (3), symétrie (4) et inégalité triangulaire (5) (qui peut jouer un rôle de pseudo-transitivité) :

(3) $D_A(X, X) = 0$

(4) $D_A(X, Y) = D_A(Y, X)$
(5) $D_A(X, Y) + D_A(Y, Z) \geq D_A(X, Z)$

Les ontologies linguistiques

Les ontologies linguistiques ont des formes arborescentes. Il s'agit d'ontologies servant à décrire le vocabulaire d'une langue. Elles sont plus particulièrement destinées à être utilisées dans une perspective de Traitement Automatique de la Langue (TAL). Le WordNet de l'Université Princeton (Felbaum, 1997) en est sans doute la plus connue des ontologies linguistiques. Il s'agit d'une base de données lexicales de l'anglais (appelée ontologie à cause de la structure hiérarchique à laquelle elle répond), qui comprend différents types d'entités lexicales (mots-composés, collocations, locutions), mais dont l'unité lexicale constitue l'unité linguistique principale de description. Sa structure n'est cependant pas organisée autour des unités lexicales individuelles. En effet, la base de données est organisée principalement à partir de la relation d'hyper-/hyponymie connectant non les unités lexicales, mais des regroupements d'unités lexicales synonymes, appelés *synsets*. Chaque *synset* fonctionne comme unité de structuration de l'ontologie, et représente plus qu'un mot lui-même ; dans une perspective cognitive, le *synset* s'approche d'une unité psycholinguistique de raisonnement. En ce sens, WordNet peut bel et bien être considéré comme une ontologie, dans la mesure où des concepts (et pas seulement des « mots ») y sont explicitement représentés.

La version 1.7 de WordNet définit ainsi le nom commun anglais *car* à l'aide de cinq *synsets* :

1. *car, auto, automobile, machine, motorcar -- (4-wheeled motor vehicle; usually propelled by an internal combustion engine; he needs a car to get to work)*
2. *car, railcar, railway car, railroad car -- (a wheeled vehicle adapted to the rails of railroad; three cars had jumped the rails)*
3. *car, gondola -- (car suspended from an airship and carrying personnel and cargo and power plant)*
4. *car, elevator car -- (where passengers ride up and down; the car was on the top floor)*
5. *cable car, car -- (a conveyance for passengers or freight on a cable railway; they took a cable car to the top of the mountain)*

Chaque *synset* dénote une acception différente du mot *car*, décrite par une courte définition. Une occurrence particulière de ce mot dénotant par exemple le premier sens (le plus courant), dans le contexte d'une phrase ou d'un énoncé, serait ainsi caractérisée par le fait qu'on pourrait remplacer le mot polysémique par l'un ou l'autre des mots du *synset sans altérer la signification de l'ensemble*.

WordNet contient également un système de catégorisation correspondant aux relations sémantiques d'hyperonymie vs. d'hyponymie comme pour le *sysnset 1 de car* :

1. car, auto, automobile, machine, motorcar
    - motor vehicle, automotive vehicle
        - vehicle
            - conveyance, transport
                - instrumentality, instrumentation

- artifact, artefact
  - object, physical object
    - entity, something

Le phénomène de la polysémie a deux aspects. Le premier aspect est qu'en dehors du cadre de la phrase, la grande majorité des prédicats verbaux traverse plusieurs catégories ontologiques *: caresser… un rêve, une personne* (*objet idéel vs. objet matériel*), *arroser… une rose, un ministre* (*végétal vs. humain),* etc. Le second aspect est que cela ne remet nullement en cause ces catégories ontologiques : dans notre vie quotidienne, nous ne confondons ni objets idéels et objets matériels, ni végétaux et humains. La pertinence de ces catégories, par ailleurs, est immédiatement reconnue dans le conflit conceptuel que l'on perçoit dans des exemples comme : **Il faut arroser les souvenirs* (Kundera, L'identité) ou **Le léger murmure semblait caresser le silence* (Revue des deux mondes).

Si la polysémie n'est pas une relation lexicale, *a fortiori*, elle ne peut pas être considérée comme un cas d'hyperonymie. Cela signifie qu'entre chaque acception d'un prédicat et l'ensemble de ses acceptions – sa polysémie – il n'y a pas une relation du type « concept particulier à concept général » (sur ce point, cf. également Kleiber 1999). Pour illustrer cette idée, faisons une liste des objets directs possibles d'un prédicat comme prendre : *prendre* un bus → monter dans ; *prendre* un boulevard → emprunter ; *prendre* un steak → commander *; prendre* une aspirine → avaler ; *prendre* un rat → capturer.

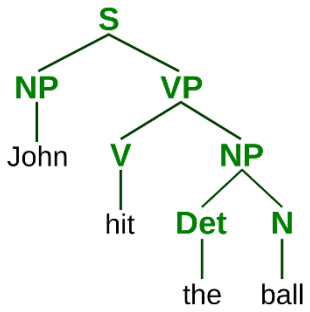
Le fait que dans toutes ces acceptions il y a le même mot -*prendre*- pousse à envisager une structure conceptuelle abstraite qui engloberait toutes les différentes acceptions. Ni les substantifs dans la deuxième colonne, ni les synonymes dans la troisième colonne ne peuvent être rapportés à un hyperonyme commun de manière évidente. Si cela est vrai, alors il est vain d'envisager un concept hyperonyme de *prendre* par rapport auquel toutes les différentes acceptions seraient des hyponymes. Chaque acception identifie bien un concept qui peut, à son tour, être hyponyme ou hyperonyme par rapport à d'autres concepts ; cependant, la polysémie de *prendre* n'est pas, elle-même, un concept fédérateur qui les rassemble.

La syntaxe

La grammaire générative de Chomsky

Le TALN s'appuie essentiellement sur des grammaires formalisables de manière à pouvoir manipuler les séquences de mots en machine. Un des premiers contributeurs a été Chomsky en deux temps :

- 1957 : Décomposition structurelle en constituants (phrase, groupes, syntagmes, mots, etc.) formalisable par des règles de réécriture. Puis règles transformationnelles (structure profonde et structure de surface)
- 1982 : Le "liage" et le "gouvernement"



Cet arbre montre la décomposition structurelle de la phrase *John hit the ball*, qui peut aussi s'obtenir par les règles suivantes :
S-> NP + VP
VP -> V + NP
NP-> Det + N
Lexique : NP-> John ; V -> hit ; Det -> the ; N-> ball

La plupart des théories et modèles utilisés en TALN (Traitement Automatique du Langage Naturel) dérivent de la théorie générativiste.

Les grammaires lexicales fonctionnelles de Bresnan-Kaplan

- **Absence de distinction entre structure profonde et structure de surface**.
- **Double structuration** :
  - Une structure de constituants (c-structure) qui est un arbre étiqueté et qui décrit directement l'agencement superficiel des éléments de la phrase et,
  - Une structure fonctionnelle (f-structure) qui distribue les notions "SUJET", "OBJET", "ADJOINT", "GENRE", "NOMBRE", etc.
- **Il n'y a pas de règles transformationnelles**.
- **Le lexique contient des informations catégorielles et fonctionnelles**. Les règles de réécriture peuvent être annotées à l'aide de "gabarits" qui permettent de colporter des informations fonctionnelles dans l'arbre syntaxique.

Ex : (a) la fille prend au bébé le jouet
(b) la fille prend le jouet au bébé

Le verbe *prendre* accepte les deux constructions :
(a) CONST prendre <(+SUJ) (+A-OBJ) (+OBJ)>
(b) CONST prendre <(+SUJ) (+OBJ) (+A-OBJ)>

et une règle unifie les deux solutions

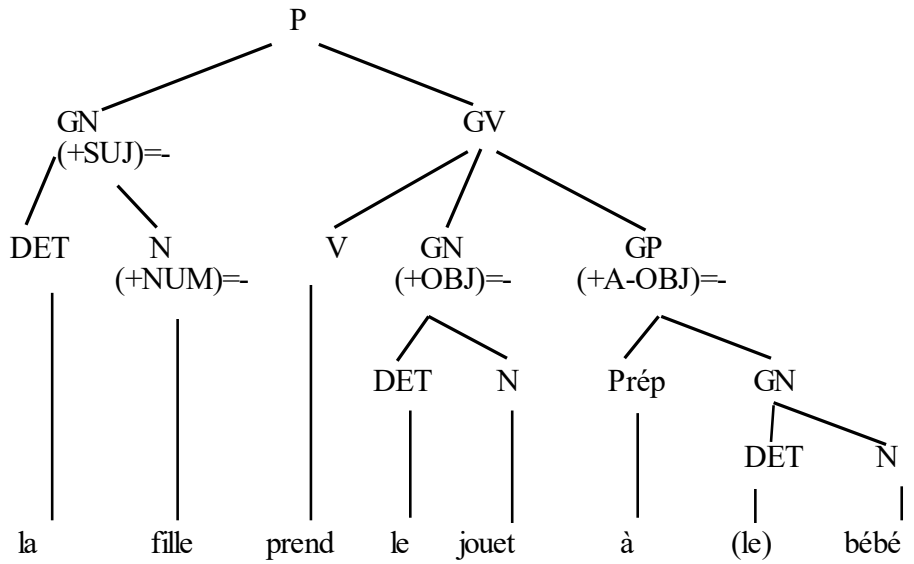
(+OBJ) <-> (+A-OBJ) (ce n'est pas une règle transformationnelle)

L'équivalence sémantique se retrouve au niveau de l'interprétation sur les notions de "agent", "bénéficiaire", "objet", etc. et non après un jeu subtil de transformations syntaxiques comme chez Chomsky. Les règles de constituants sont pour cet exemple :

P -> GN . GV /(+SUJ)=-, +=-/
GN -> DET . N /(), (+NUM)=-/
GV -> V . GN . GP /(), (+OBJ)=-, (+A-OBJ)=-/
GP -> Prép . GN

+ et - notent les relations de dominance sur le père (+) et les fils (-) qui sont à instancier au niveau de la structure fonctionnelle. SUJ, OBJ etc. sont des métavariables.

Représentation de l'arbre syntaxique "décoré" :

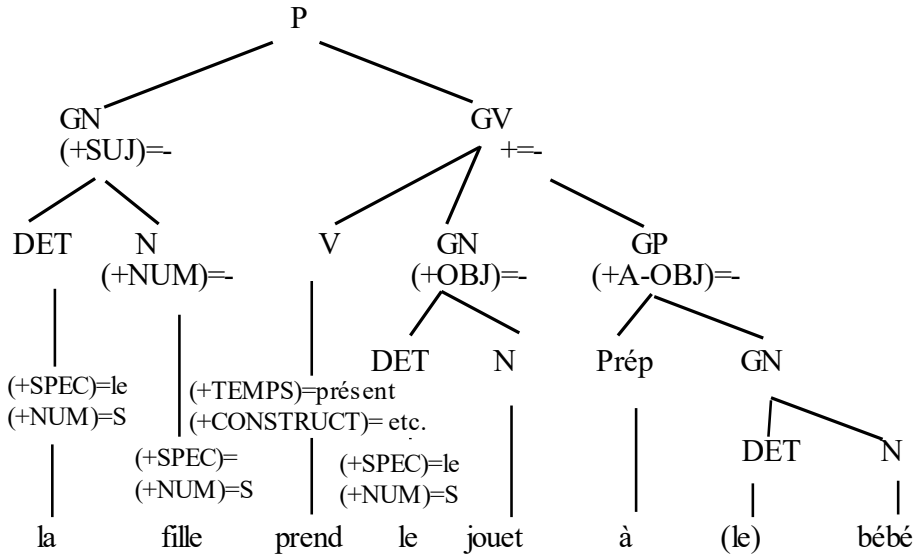


Représentation du lexique :

la	DET, (+SPEC)=le, (+GENRE)=F, (+NUM)=S
fille	N, (+SEM)="Personne", (+GENRE)=F, (+NUM)=S
prendre	V, (+TEMPS)="Présent", (+MODE)="Ind", (+PERS)=3, (+NUM)=S, CONSTRUCT
prendre	<(+SUJ) (+A-OBJ) (+OBJ)>
le	DET, (+SPEC)=le, (+GENRE)=M, (+NUM)=S
jouet	N, (+SEM)="Objet", (+GENRE)=M, (+NUM)=S
au	DET, (+SPEC)=à le, (+GENRE)=M, (+NUM)=S
bébé	N, (+SEM)="Personne", (+GENRE)=M, (+NUM)=S

L'analyse se fait en propageant les valeurs des métavariables jusqu'au lexique

(ascendant ou descendant selon les indications fournies par les opérateurs + et -), puis en remontant le long des nœuds, on s'assure de la cohérence des attributs.



Ce qui donne l'analyse suivante :

```
[SUJ=fille
        SPEC=le
        NUM=singulier
        GENRE=féminin
        SEM=personne
ACTION=prendre
        TEMPS=présent
        MODE=indicatif
        PERS=3
        NUM=singulier
        CONSTRUCT prendre <(+SUJ) (+A-OBJ) (+OBJ)>
OBJ=jouet
        SPEC=le
        NUM=singulier
        GENRE=masculin
        SEM=objet
A-OBJ=bébé
        SPEC=à le
        NUM=singulier
        GENRE=masculin
        SEM=personne]
```

qui est appelée la **structure fonctionnelle** de la phrase. On remarque le rôle central du

verbe avec son cadre rectionnel.

Il est difficile dans ce cadre d'élaborer une syntaxe des adverbiaux et des subordonnées circonstancielles. Ce formalisme fonctionne bien pour des relations de dominance proche. Dans un autre cas il faut introduire des métavariables ayant une portée plus étendue (++ et --)

Les grammaires catégorielles (GPSG) de Gazdar et grammaires dérivés (HPSG, TAG)

Ces grammaires sont fondées sur le lambda-calcul et la notion de catégories —qui héritent les unes des autres.
Exemple, la catégorie SN

D2

$$\begin{bmatrix} \text{Cat} = \text{SN} \\ \text{agr} = \begin{bmatrix} \text{nombre} = \text{sing} \\ \text{personne} = 3 \end{bmatrix} \\ \text{subj} = \langle \text{Indices} \rangle \end{bmatrix}$$

agr et subj sont des traits. Parmi ceux-ci le trait BAR joue un rôle prépondérant, il donne le degré de profondeur dans la structure arborescente finale.

Gazdar remarque qu'une règle de la forme SV -> V SN SP est ambigüe dans la mesure où elle ne permet pas de distinguer la notion de dominance de celle de succession. Il propose de les distinguer par l'écriture :

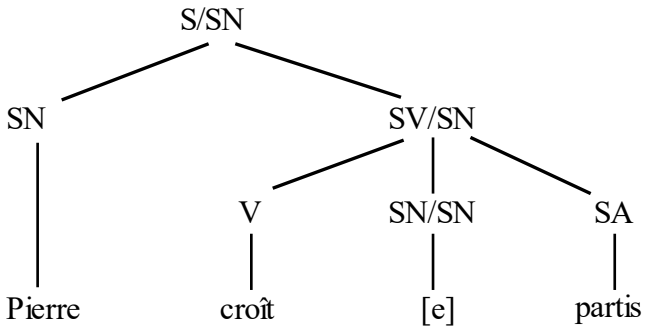
SV -> V SN SP (domination)
[SubCat] < ~[SubCat] (succession) où SubCat est un trait de sous-catégorisation

Il obtient également une économie de règles en propageant les traits par héritage dans les catégories.

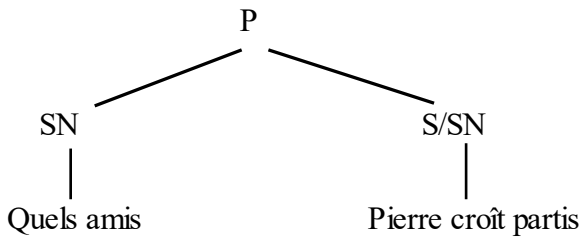
La syntaxe d'une phrase peut donc être représentée par un arbre construit à partir de sous-arbres : la grammaire se réduit à des règles de formation d'arbres (imbrication de sous-arbres en utilisant le principe de tête, de pied, d'accord, c'est-à-dire les relations entre les sous-arbres). Il définit aussi des métarègles pour déduire d'autres règles de formation —ex: pour former le passif RègleActif -> RèglePassif.

Soit la phrase : Quels amis, Pierre croit partis ?

Elle est d'abord traitée par une grammaire transformationnelle pour rapprocher les termes qui doivent s'accorder, Pierre croit ((Quels amis) partis ?) qui peut être analysée comme une phrase à laquelle il manque un SN, c'est-à-dire P/SN. Le "/" peut être lu comme "manque" et l'arbre comme une propagation de ce "manque".



Complété par la règle de topicalisation : SP -> SN P/SN



Les principes d'instanciation :

1. HFC = Head Feature Convention, les traits de tête doivent avoir la même valeur pour le père et le fils tête
2. FFP = Foot Feature Principle, idem 1. pour le père et tous ses fils
3. Accord entre les traits des frères

Forme logique de Montague

C'est une composante sémantique logique associée à la GPSG. Les idées principales sont :
1. Application de la notion de la vérité dans l'interprétation des énoncés.
2. Déclaration de la variété des mondes possibles, dans lesquels on interprète un énoncé.
3. Acceptation du Principe de Compositionnalité.
4. Utilisation de types sémantiques (catégories d'après Montague)
5. Uniformité structurale des règles syntaxiques et sémantiques.
6. Utilisation d'un langage formel pour décrire des phénomènes linguistiques :
    6.1. Calcul des prédicats 6.2. Lambda-calcul

Formellement, les types sont définis de la manière suivante :
Types sémantiques
    1) e et t sont des types sémantiques
    2) Si a et b sont des types sémantiques, alors est aussi un type sémantique
    3) Rien d'autre n'est un type sémantique
Domaines des dénotations sémantiques
    1) De := D (l'ensemble des individus)
    2) Dt := {0, 1} (l'ensemble des valeurs de vérité)

3) Pour tous les types sémantiques a et b, D est l'ensemble de toutes les fonctions de Da vers Db

Cette définition est récursive, elle « génère » un ensemble infini de types, dont seulement une petite partie est utilisée en description d'une langue.

Exemple : analyse de la phrase *Chaque homme aime une femme*

Chaque homme : DET + NOUN → np

[(λQ.(λP.(∀x.(Q(x)>P(x)))))@(homme(y))] =

= [(λP.(∀x.(homme(x)>P(x))))] =α

une femme : DET + NOUN → np

[(λQ.(λP.∃y(Q(y)&P(y))))@(femme(z))] =

= [(λP.∃y.(femme(y)&P(y)))] = β

aime [une femme] : VERB + np → vp

[(λK.(λx.(K@(λy.(aimer(y,x))))))@(β)] =

= [(λK.(λx.(K@(λy.(aimer(y,x))))))@([(λP.∃z.(femme(z)&P(z)))])] =

= [(λx.(λP.∃z.(femme(z)&P(z)))@(λy.(aimer(y,x)))) =

= [(λx.(∃z.(femme(z)&aimer(y,x)))] = γ

[Chaque homme] [aime [une femme]] : dp + vp → S

[α@(γ)] = [([(λP.(∀x.(homme(x)>P(x))))])@(γ)] =

= [(λP.(∀x.(homme(x)>P(x))))@(λr.(∃z.(femme(z)&aimer(y,r)))] = =

[(∀x.(homme(x)>(λr.(∃z.(femme(z)&aimer(z,r))))))@(x)] =

= [(∀x.(homme(x)>(∃z.(femme(z)&aimer(z,x)))))]

La formule finale est : (∀x.(homme(x)>(∃z.(femme(z)&aimer(z,x)))))

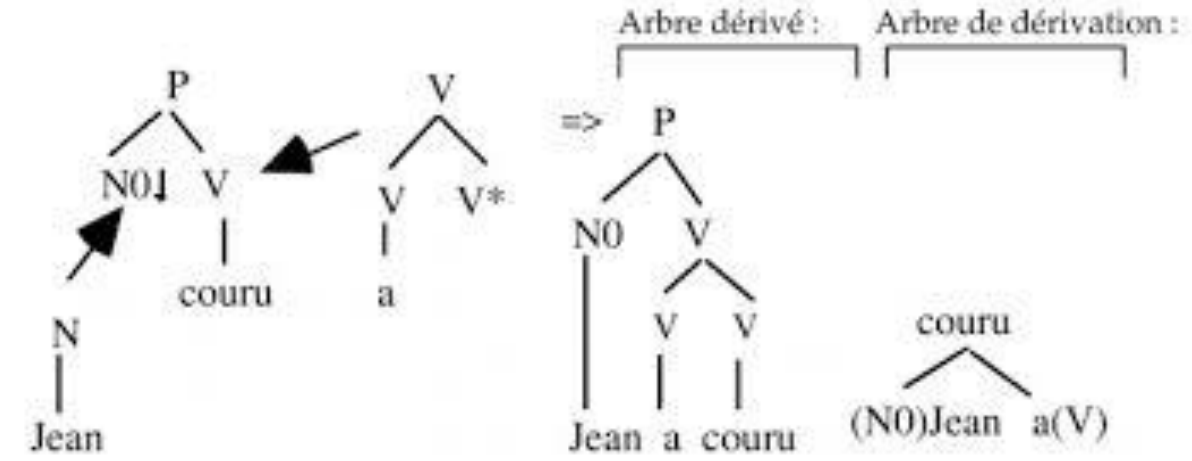
Les grammaires d'arbres adjoints (TAG)

Les arbres élémentaires :

Une grammaire TAG, définie par Joshi [Joshi, 1975], est un ensemble fini d'arbres élémentaires qui peuvent être de deux types : arbres initiaux (notés α) ou arbres auxiliaires (notés β). Ces arbres, dont la profondeur peut être supérieure à 1, ont à leurs feuilles des terminaux ou des nœuds à substitution. Les arbres auxiliaires ont en outre un nœud feuille (appelé nœud « pied ») étiqueté par un non terminal de même catégorie que celle de leur nœud racine. La combinaison de plusieurs arbres élémentaires produit des arbres « dérivés » (notés γ) dont les feuilles définissent les séquences de terminaux appartenant au langage d'une grammaire TAG.

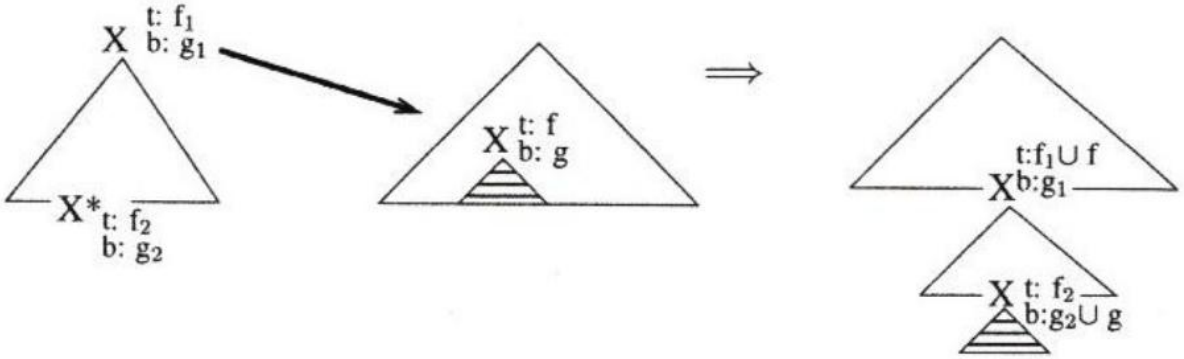
L'adjonction et la substitution :

L'opération de combinaison principale est l'adjonction, qui insère un arbre auxiliaire (ou dérivé d'un arbre auxiliaire) à un nœud quelconque de même catégorie dans un arbre élémentaire ou dérivé. L'opération de substitution, qui est similaire à l'opération de concaténation utilisée dans les grammaires hors contexte, insère un arbre initial (ou dérivé d'un arbre initial) à la frontière d'un arbre élémentaire ou dérivé.

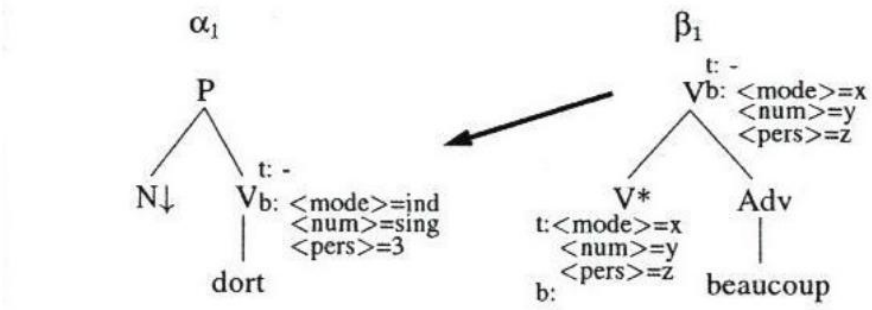


Les traits et l'unification :

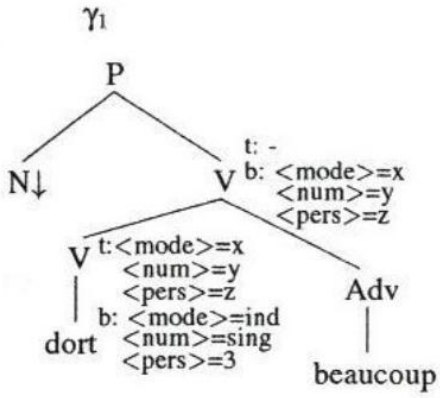
Les traits sont des couples attribut-valeur, les valeurs pouvant être des symboles atomiques ou des traits. Des ensembles de traits (conjoints) peuvent être associés à un mot ou à un syntagme. L'unification entre deux structures de traits produit une structure résultante sauf si les deux structures portent des informations incompatibles (par exemple, si un attribut présent dans les deux structures a une valeur différente) ; on dit alors que l'unification « échoue ». La structure résultante est la plus petite structure qui contient toute l'information contenue dans la première structure et toute l'information contenue dans la deuxième structure.



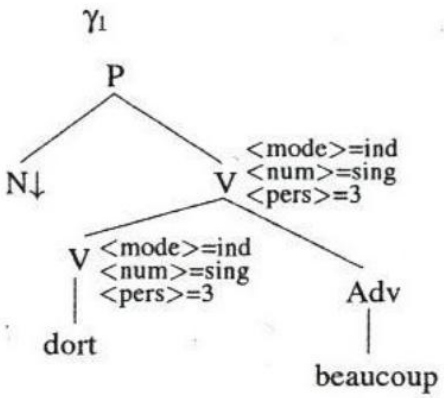
Pour les grammaires TAG, la mise à jour des structures de traits a été définie pour tenir compte de cette opération particulière qu'est l'adjonction. A chaque arbre élémentaire est associé au départ une structure en traits bipartite, divisée en une partie « amont » (en anglais « top ») et une partie « aval » (en anglais « bottom »). En « amont » sont les traits indiquant les relations du nœud avec les nœuds qui le dominent, en « aval », les relations avec ceux qu'il domine. Les nœuds auxquels ne peut se produire aucune adjonction, par exemple les nœuds à substitution et les nœuds pieds des arbres auxiliaires, peuvent avoir une structure de traits unique. A la fin d'une dérivation, parties amont et aval doivent s'unifier à chaque nœud. A cause de l'opération d'adjonction, l'unification a donc lieu en deux temps dans une dérivation : entre les structures de traits des arbres élémentaires combinés au fur et à mesure, puis à chacun des nœuds de l'arbre dérivé ainsi obtenu lorsque toutes les combinaisons ont été effectuées.



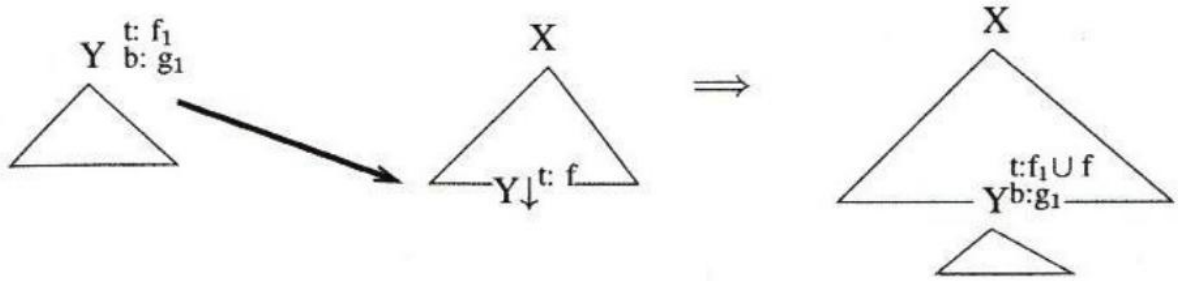
après adjonction :

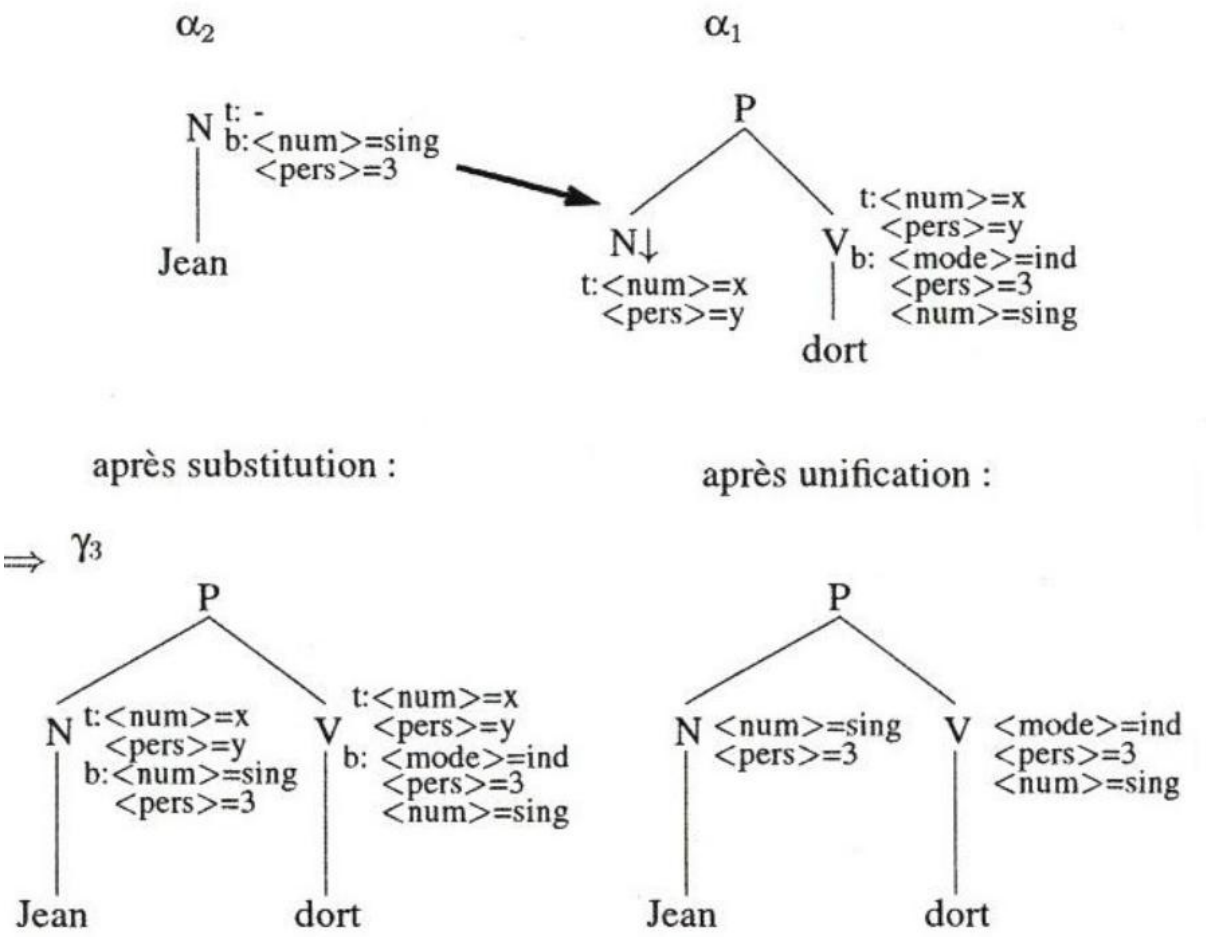


après unification :



Pour qu'il y ait substitution, les traits amont du nœud racine de l'arbre substitué doivent s'unifier avec les traits (amont) du nœud où a lieu la substitution :





Les grammaires cognitives

On regroupe sous l'appellation « grammaires cognitives » un courant de recherche en linguistique qui est né dans les années 1980, sur la côte Ouest des Etats-Unis. Ce courant a rapidement acquis une large audience internationale, grâce notamment aux textes fondateurs de quatre auteurs : Ronald Langacker, avec le premier tome de *Foundations of Cognitive Grammar* (Langacker 1987) ; Leonard Talmy, avec deux articles essentiels : « Lorce Dynamics in Language and Thought » et « The Relation of Grammar to Cognition » (Talmy *1988a* et *1988b*), articles repris dans l'ouvrage récent *Towards a Cognitive Semantics* (Talmy 2000) ; Georges Lakoff, avec *Women, Fire and Dangerous Things* (Lakoff 1987) ; et enfin Gilles Fauconnier, avec un ouvrage écrit d'abord en français, *Espaces mentaux* (Fauconnier 1984), aussitôt traduit en anglais (Fauconnier 1985), et réédité par la suite avec une nouvelle préface (Fauconnier 1994).

Les fondements théoriques des grammaires cognitives reposent sur quelques principes, lesquels donnent à ce courant son unité et son originalité. En premier lieu, l'activité de langage, tout en ayant ses spécificités, doit être régie par des *mécanismes cognitifs généraux,* à l'œuvre dans toutes les activités cognitives. Ainsi, par exemple, comme nous aurons l'occasion de le voir plus en détail, l'opposition gestaltiste entre figure et fond se retrouve dans l'organisation des énoncés linguistiques. Plus généralement, la perception visuelle et l'expérience sensori-motrice jouent un rôle central dans la compréhension de la structure sémantique du langage.

En conséquence, les grammaires cognitives rejettent totalement la primauté et l'autonomie accordées par les grammaires génératives à la syntaxe. L'étude des structures syntaxiques n'est pas une finalité en soi, qui permettrait de découvrir l'essence même du langage. Au contraire, les constructions syntaxiques sont, au même titre que les autres éléments constitutifs des langues (les unités lexicales et grammaticales), des structures symboliques, porteuses de sens, qui contribuent à la signification globale des énoncés.

C'est donc la *sémantique* qui est placée au cœur du dispositif. La finalité du langage est de construire des structures sémantiques complexes, que Talmy appelle « représentations cognitives », Langacker « structures conceptuelles » et Fauconnier « espaces mentaux ». L'étude de la grammaire consiste à rendre compte de la manière dont les unités linguistiques, sortes de « briques » élémentaires symboliques, se combinent pour produire des représentations complexes. Chaque différence de forme correspond à des différences dans la représentation construite. Ainsi, pour Langacker, les deux énoncés suivants n'ont pas le même sens :

(1) He sent a letter to Susan
(2) He sent Susan a letter

Même s'ils décrivent le même événement, ils ne le présentent pas de la même manière : l'énoncé (1), à cause de la préposition *to,* met en relief la trajectoire de la lettre, alors que l'énoncé (2) met l'accent sur le résultat de l'action, la possession de la lettre par Susan. Ces différences de focalisation (de « profilage », dans la terminologie de Langacker) doivent faire partie intégrante de la description sémantique de ces deux énoncés. Deux paraphrases, aussi proches soient-elles, n'ont donc pas la même représentation sémantique associée.

Les modèles de langage (LLM) probabilistes

L'intelligence artificielle et ses sous-disciplines connaissent un engouement croissant, notamment depuis les résultats impressionnants obtenus par les réseaux de neurones profonds. Le traitement automatique du langage naturel, qui est une sous-discipline de l'intelligence artificielle, a ainsi franchi une étape importante grâce à ces modèles. L'intérêt croissant pour le traitement automatique du langage naturel s'explique par le large éventail d'applications dans l'industrie, notamment dans les domaines de la traduction automatique, de la classification de textes, du résumé automatique, de la reconnaissance des entités nommées, de l'analyse des sentiments et des agents conversationnels.

Word2Vec utilise une technique de fenêtrage local pour déterminer le mot central et les mots voisins qui représentent le contexte de ce mot. Il utilise un simple perceptron pour apprendre à représenter les mots dans deux matrices formées à partir des poids du perceptron. La première contient les mots centraux et la deuxième, les mots du contexte. Le modèle SkipGram prédit les mots du contexte à partir du mot central, tandis que CBoW prédit le mot central à partir des mots du contexte.

Modèle Skip-gram : le modèle Skip-gram maximise la probabilité de prédire le contexte

C autour d'un mot cible w_t. La fonction objective est donnée par :

$$max \prod_{t=1}^{T} \prod_{c \in C_t} P(w_c|w_t)$$

Où la probabilité conditionnelle est définie comme :

$$P(w_c|w_t) = \frac{e^u}{\sum_{w \in V} e^u}$$

Avec : $u = v'_{wc} - v_{wt}$ et v_{wt} vecteur d'entrée du mot cible w_t, ; v'_{wc} est le vecteur de sortie du mot de contexte w_c, ; V est le vocabulaire.

Modèle CBOW (Continuous Bag of Words) : le modèle CBOW prédit un mot cible w_t à partir des mots de contexte C environnants. La fonction objective est donnée par :

$$max \prod_{t=1}^{T} P(w_t|C_t)$$

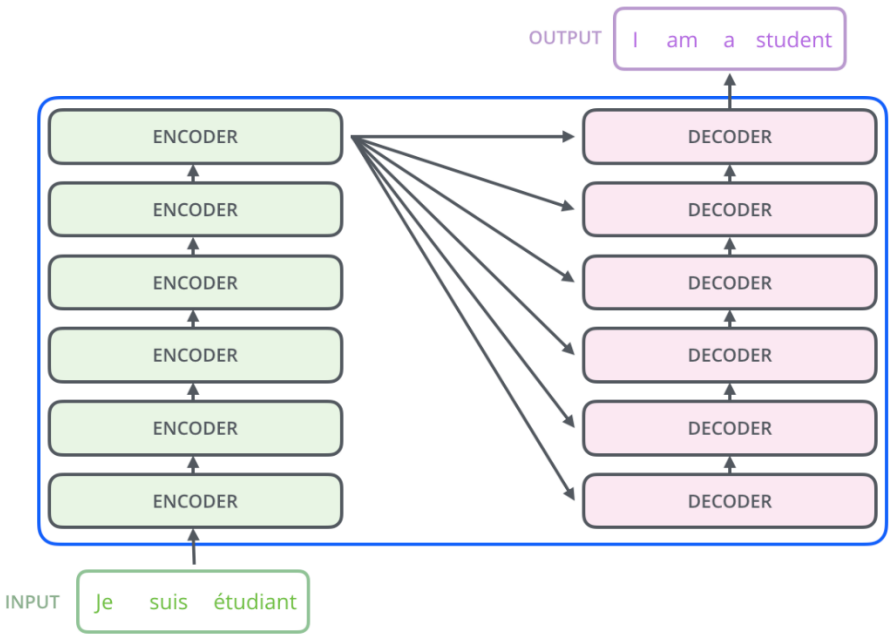
Où la probabilité conditionnelle est :

$$P(w_t|C_t) = \frac{e^u}{\sum_{w \in V} e^u}$$

Avec : $u = v'_{wt}$ - moy des vecteurs de contexte v_{wc}

Principe du réseau Transformer

L'architecture **Transformer**, introduite par l'article *Attention Is All You Need* en 2017, a révolutionné l'IA en abandonnant les structures récurrentes (RNN) au profit d'un mécanisme de "self-attention". Voici une décomposition de ses modules et de son fonctionnement.

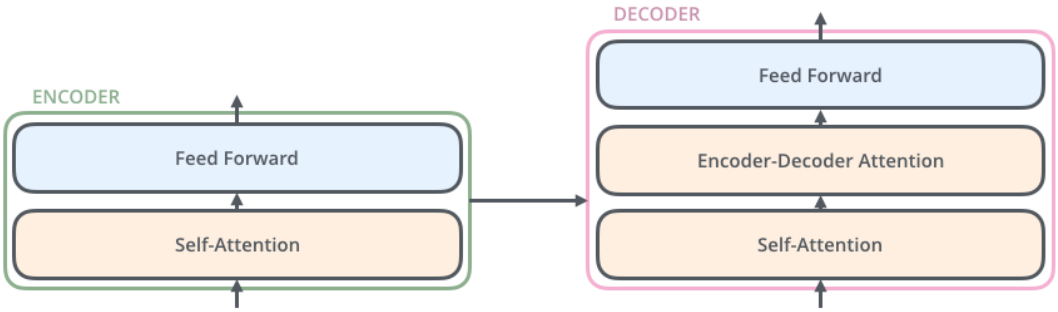


1. L'Architecture Globale : Encodeur-Décodeur

Le Transformer original repose sur une structure symétrique, bien que de nombreux

modèles modernes (comme GPT) n'utilisent que la partie décodeur.

- **L'Encodeur :** Il traite la séquence d'entrée (par exemple, une phrase en français) pour en extraire une représentation mathématique riche en contexte.
- **Le Décodeur :** Il utilise les informations de l'encodeur pour générer une séquence de sortie (par exemple, la traduction en anglais), un mot à la fois.



2. Les Modules Clés du Transformer

A. L'Input Embedding et le Positional Encoding

Comme le Transformer traite tous les mots d'une phrase en même temps (parallélisation), il ne sait pas naturellement quel mot vient avant l'autre.

- **Embedding :** Transforme les mots en vecteurs de nombres.
- **Positional Encoding :** Ajoute un signal mathématique (souvent via des fonctions sinus et cosinus) à chaque vecteur pour indiquer sa position dans la phrase.

$e_t = E(x_t) + p_t$

$E(x_t) \in \mathbb{R}^d$: embedding lexical ; $p_t \in \mathbb{R}^d$: encodage positionnel

B. Le mécanisme de Multi-Head Attention (MHA)

C'est le cœur du modèle. Il permet au réseau de focaliser son attention sur les mots les plus pertinents par rapport au mot traité.

- **Self-Attention :** Calcule les relations entre chaque mot de la même phrase.
- **Multi-Head :** Le modèle effectue ce calcul plusieurs fois en parallèle pour capturer différentes nuances (ex : une "tête" surveille la grammaire, une autre le sens des verbes, etc.).

Soit $H^{(l)} \in \mathbb{R}^{T\times d}$ la représentation à la couche l. On calcule à l'aide des matrices W apprises :

$Q = H^{(l)}W_Q$; $K = H^{(l)}W_K$; $V = H^{(l)}W_V$

— Q ∈ $R^{n\times dk}$ est la matrice des requêtes

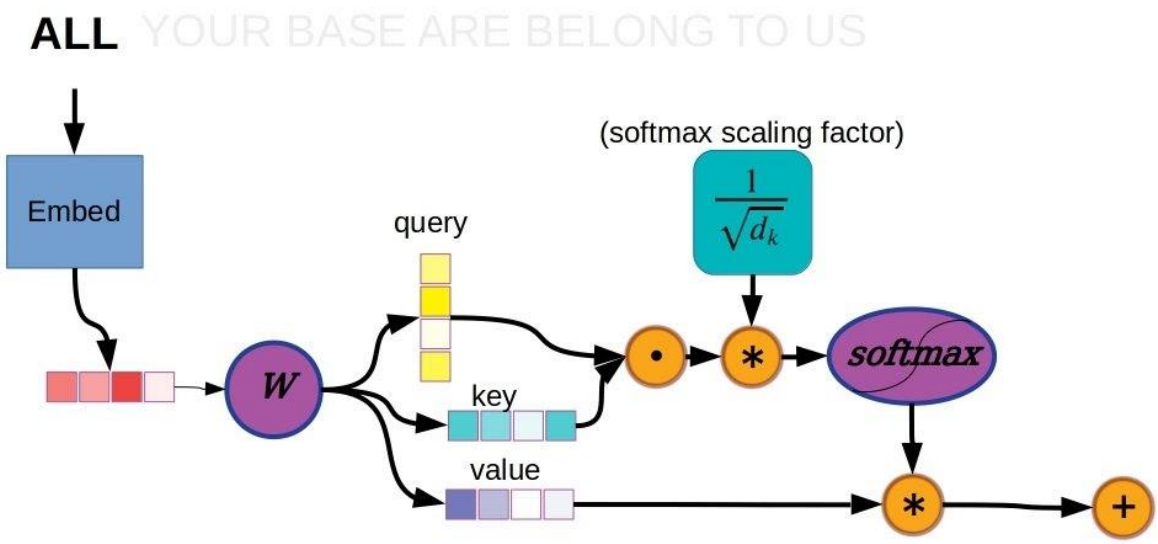
— K ∈ $R^{n\times dk}$ est la matrice des clés

— V ∈ $R^{n\times dv}$ est la matrice des valeurs

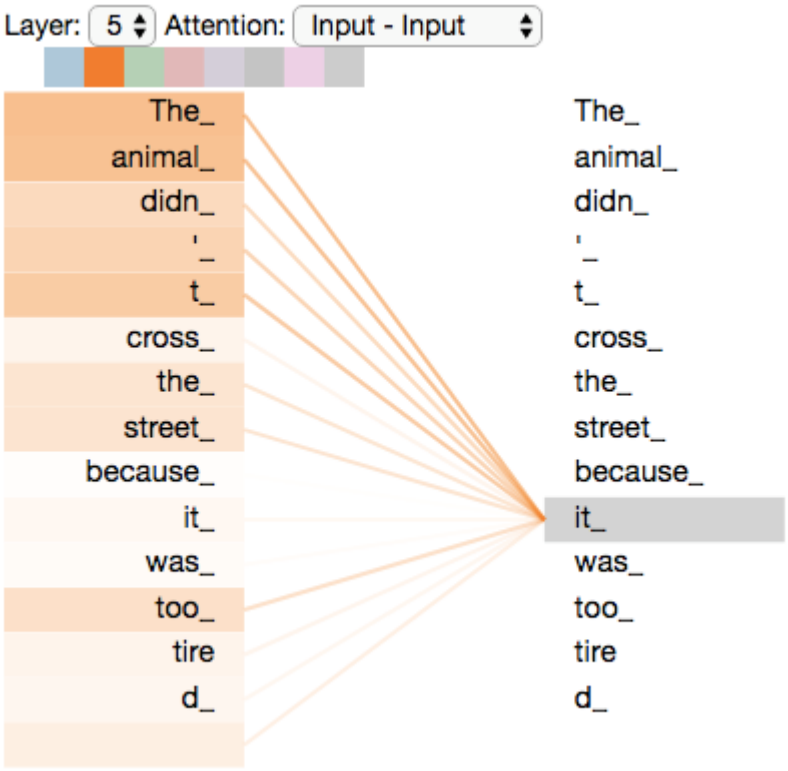
— d_k est la dimension des clés.

Attention : $\text{Att}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M\right)V$

Avec M= matrice de masque causal (empêche l'accès au futur)



Multi-head : $\text{MHA}(H) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W_O$
Exemple de self attention de « it » sur les mots de l'énoncé



C. Le Feed-Forward Network (FFN)
Après l'attention, chaque mot passe par un petit réseau de neurones classique (entièrement connecté). Cela permet de traiter les informations extraites par l'attention de manière plus profonde.
$\text{FFN}(x) = \sigma(xW_1 + b_1)W_2 + b_2$

$H^{(l+1)} = \text{LayerNorm}(\widetilde{H} + \text{FFN}(\widetilde{H}))$
Soit après L couches : $H^{(L)}$

D. Normalisation et Connexions Résiduelles (Add & Norm)
Pour éviter que le signal ne se perde ou ne sature pendant l'entraînement, chaque sous-couche est entourée d'une connexion résiduelle suivie d'une normalisation de couche (**Layer Normalization**).
$\widetilde{H} = \text{LayerNorm}(H^{(l)} + \text{MHA}(H^{(l)}))$

E. Projection finale
$z_t = H_t^{(L)} W_{vocab}$
$P_\theta(x_t \mid x_{<t}) = \text{softmax}(z_t)$

Le flux de données en résumé

Étape	Module	Rôle
1	**Entrée**	Conversion du texte en vecteurs + ajout de la position.
2	**Attention**	Compréhension des relations entre les mots.
3	**FFN**	Transformation individuelle des vecteurs de mots.
4	**Sortie**	Couche *Linear* et *Softmax* pour prédire le mot suivant.

L'architecture Transformeur est utilisée dans de nombreux LLM notamment par BERT (Bidirectional Encoder Representations from Transformers) qui est un modèle de traitement automatique du langage développé par Google en 2018.
BERT est pré-entraîné sur de grandes quantités de textes à l'aide de deux tâches principales :

- **Masked Language Modeling (MLM)** : prédire des mots masqués dans une phrase.
- **Next Sentence Prediction (NSP)** : déterminer si une phrase suit logiquement une autre.

Après pré-entraînement, il peut être **ajusté (fine-tuning)** pour de nombreuses tâches : classification de texte, réponse à des questions, reconnaissance d'entités nommées, etc.
BERT a marqué un tournant majeur en TALN en améliorant significativement les performances sur de nombreux benchmarks et en popularisant l'usage des modèles pré-entraînés.

Principe de chatGPT

1. L'architecture : Complet vs Décodeur seul

- **Le Transformer (Original) :** Il possède deux bras (un **Encodeur** pour comprendre et un **Décodeur** pour générer). Il a été conçu pour la traduction (ex : Français → Anglais).
- **ChatGPT (basé sur GPT) :** GPT signifie *Generative Pre-trained Transformer*. Il utilise uniquement la partie **Décodeur**. Au lieu de traduire, il se contente de "prédire le mot suivant" de façon extrêmement sophistiquée.

2. L'entraînement : L'étape de "l'éducation"
Le Transformer est une architecture vide. ChatGPT a subi un entraînement massif en trois étapes :

1. **Pré-entraînement :** Il a lu une immense partie d'Internet pour apprendre la

grammaire et les faits.

2. **Fine-tuning :** On l'a spécialisé pour répondre à des instructions (pour qu'il ne se contente pas de compléter des phrases, mais qu'il réponde à des questions).
3. **RLHF (Reinforcement Learning from Human Feedback) :** Des humains ont noté ses réponses pour lui apprendre à être poli, utile et à éviter les contenus dangereux.

L'usage : Outil technique vs Interface conversationnelle

Caractéristique	Transformer	ChatGPT
Nature	Algorithme / Architecture	Application / Modèle de langage
Entrée	Données brutes (vecteurs)	Texte naturel (Prompts)
Capacité	Peut tout faire (vision, texte, son)	Optimisé pour le dialogue et le code

La prédiction du mot suivant est le cœur du fonctionnement de ChatGPT. Contrairement à un humain qui a une pensée globale, le modèle traite le langage comme un calcul de probabilités. Voici comment cela se passe, étape par étape :

1. La découpe en "Tokens"
ChatGPT ne lit pas les mots entiers, mais des morceaux de mots appelés **tokens**.

- Un mot court comme "chat" est 1 token.
- Un mot long ou complexe comme "anticonstitutionnellement" peut être divisé en 5 ou 6 tokens.

En moyenne, 1 000 tokens correspondent à environ 750 mots.

2. Le calcul des probabilités
Lorsque vous écrivez une phrase, le modèle analyse toute la séquence précédente pour calculer quel est le token le plus probable pour la suite.
Imaginons que je doive compléter : **"Le ciel est..."** Le modèle génère une liste de probabilités : **bleu** : 85% ; **gris** : 10% ; **nuageux** : 4% ; **vert** : 1%

3. Le mécanisme de "Sampling" (L'échantillonnage)
Si le modèle choisissait **toujours** le mot le plus probable (1er de la liste), il serait très répétitif et monotone. Pour le rendre créatif et naturel, on utilise des paramètres :

- **La Température :** Si elle est basse (0.2), le modèle est très prudent et choisit presque toujours le mot le plus probable. Si elle est haute (0.8+), il prend des risques et choisit parfois des mots moins probables, ce qui le rend plus "créatif".
- **Top-k :** Le modèle ne regarde que les mots dont la probabilité cumulée atteint un certain seuil (ou les k premiers), éliminant ainsi les choix absurdes (comme "vert" dans l'exemple du ciel).

4. Une boucle infinie (Autorégressif)
C'est l'aspect le plus important : le modèle est **autorégressif**.

1. Il prédit le mot A.
2. Il ajoute le mot A à la phrase initiale.
3. Il recommence pour prédire le mot B, et ainsi de suite.

Initialisation avec un prompt $x_{1:k}$

Pour $t > k$: $x_t \sim P_\theta(\cdot | x_{<t})$
Selon la stratégie : argmax (greedy), sampling, top-p, nucleus (top-p)

5. Alignement par RLHF
On introduit un modèle de récompense : $R_\phi(x, y)$
où : x= prompt ; y= réponse générée
Objectif PPO (simplifié) :
$\max_\theta \ \mathbb{E}_{y\sim P_\theta(\cdot|x)}\left[R_\phi(x,y) - \beta\, \mathrm{KL}(P_\theta \,||\, P_{\text{base}})\right]$
Cela contraint le modèle à : maximiser la récompense humaine tout en restant proche du modèle de base

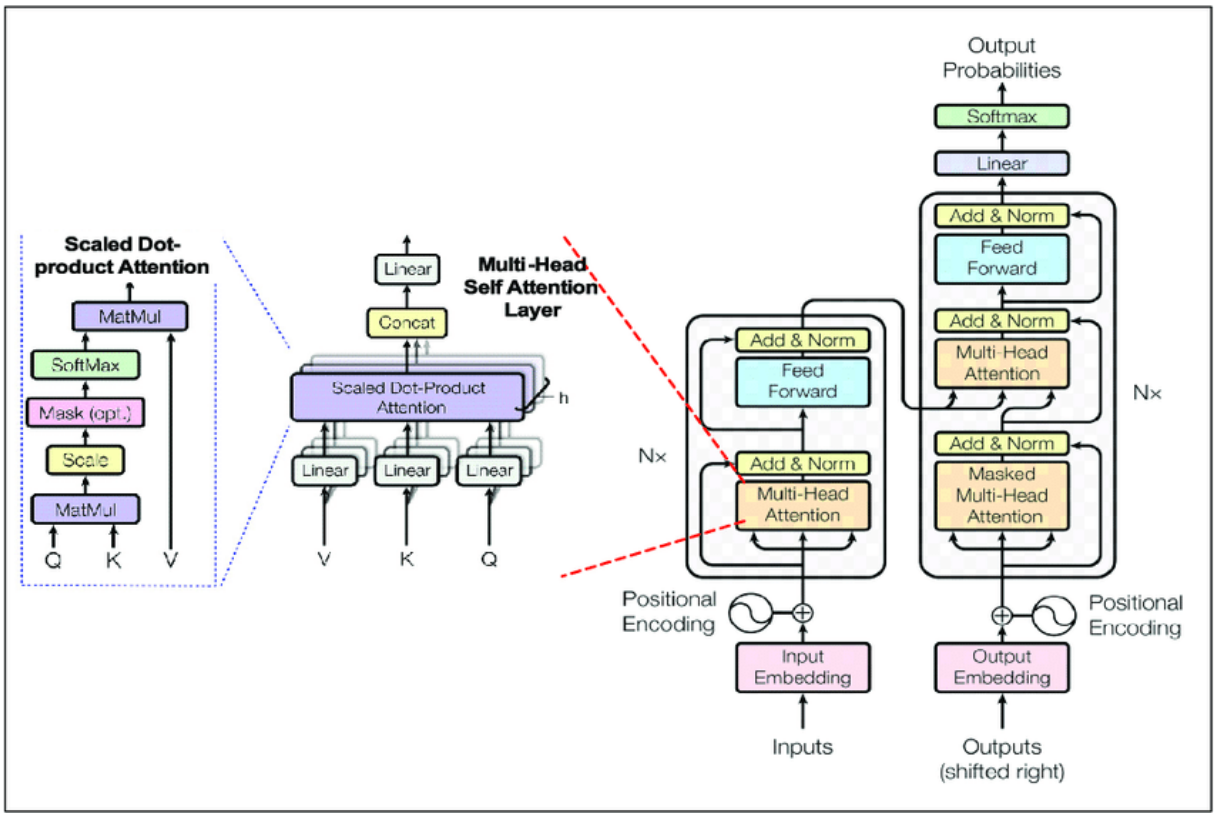
Résumé
On peut résumer ChatGPT comme le système :

$$\theta^* = \arg\max_\theta \underbrace{(\mathbb{E}[\log P_\theta(x)]}_{\text{pré-entraînement}} + \underbrace{\mathbb{E}[R_\phi(x,y)]}_{\text{alignement}} - \beta\, \mathrm{KL})$$

avec génération : $y = \mathrm{Decode}(P_{\theta^*}(\cdot | x))$

Problème des hallucinations)
Comme ChatGPT ne fait que "calculer le mot le plus probable statistiquement", il n'a pas de notion de **vérité**. S'il prédit une suite de mots qui a l'air très convaincante grammaticalement mais qui est fausse sémantiquement c'est parce que statistiquement, ces mots "vont bien ensemble" dans sa base de données.

Schéma général BERT et GPT



Les modèles de langue d'aujourd'hui ne se limitent pas au traitement purement symbolique des textes. Ils prennent désormais en compte les structures syntaxiques, mais aussi visuelles, afin de mieux encoder le sens des mots. L'intégration explicite de la syntaxe permet de mieux comprendre comment les modèles de langue génèrent des représentations numériques des mots. Les modèles multimodaux, qui combinent texte et image, apportent une nouvelle dimension à l'encodage du sens des mots. L'arrimage d'images et de mots permet à ces modèles de dépasser la simple représentation textuelle et d'ancrer le monde réel dans la langue. Cela permet d'offrir davantage d'explicabilité et une vision plus complète de la manière dont les grands modèles de langue encodent le sens des mots.

La sémantique

La sémantique descriptive

La sémantique descriptive s'attache à décrire le sens des mots

Antonymie : contradictoire : beau/laid
Complémentaire : acheter/vendre
Graduelle : brûlant/chaud/tiède/froid/glacial
Synonymie : paradigmatique : abdomen/ventre/bidon
Syntagmatique : connaître le succés/être réputé
Homonymie : homographique et/ou homophonique
Louer (une voiture)/(quelqu'un=faire des louanges)
conter/compter
Polysémie : plusieurs sens : lettre=caractère/missive/culture

Formes de changement :

Ellipse : mon général=substantivisation de capitaine général
Métaphore : analogie lunettes=origine : lune
Métonymie : métaphore de contiguité: casier judiciaire
Synecdoque : figure de sens (trope) qui réduit le sens d'un objet à l'une de ses propriétés : allez les verts
Extension : panier=origine : pour le pain
Restriction : sens commun->sens spécialisé : maraudage (juridique)
Anaphore : référence indirecte rétrograde= cet exemple, prends-le
Cataphore : -id- mais progrédient= quand il entra, le facteur...

Représentation des sèmes :

Table d'attributs ou traits

	A	B	C
geai	+	+	+
pigeon	-	+	-

A: passereau
B: taille voisine de 35 cm
C: plumage bigarré

Prédicats

x='geai'
y='pigeon'
taille(x)=35cm ± ε, oiseau(x)=passereau ET taille(x)=taille(y) ET plumage(x)=bigarré
/pigeon, passereau, bigarré/ sont des prototypes ou des propriétés

Frames

passereau: sorte_de oiseau
taille: [pigeon, hibou, aigle]
plumage: [gris, strié, bigarré]
geai: est_un passereau
taille=pigeon
plumage=bigarré

La sémantique générative

Le problème de la compétence/performance se pose ici en clair puisque syntaxe et sémantique sont réduites aux mêmes représentations (structures parenthésées ou arborescentes déduites de règles de réécriture)

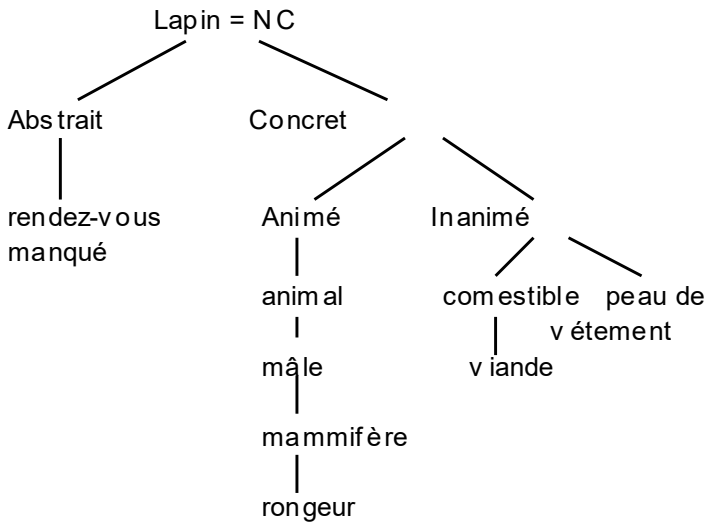
Les ambiguïtés peuvent être levées (a) par le rapport qu'entretiennent certains mots dans la phrase ou le texte (b) à l'aide d'attributs incompatibles :

(a) la note est juste mais un peu chère ==> note au sens de "addition"
(b) la peinture est silencieuse ==> (?)

La sous-catégorisation syntaxique trop poussée fait apparaitre les distinctions sémantiques (Chomsky est partisan de cette approche). Il faut donc séparer les rôles syntaxe/sémantique dans le lexique. Les entrées de ce dernier sont:

- catégories syntaxiques
- traits sémiques (ou attributs)
- différentiateurs (indiquent les différents sens selon le contexte)

Ex : Lexique sémantique



Les différentiateurs permettent de s'aiguiller dans le graphe (on peut aussi dupliquer les entrées lexicales) comme suit :

1. Lapin [NC, Abstrait, rendez-vous]

2. Lapin [NC, Concret, Animé, animal, mâle, mammifère, rongeur]
3. Lapin [NC, Concret, inanimé, comestible, viande]
4. Lapin [NC, Concret, inanimé, peau])

Analyse par amalgame et projection

La composante d'interprétation sémantique est : la manière dont les significations se combinent qui est fixée par la structure syntagmatique de la phrase : les règles associeront d'abord les sens des unités dominées par les nœuds les plus bas puis rassembleront les informations ainsi obtenues au niveau des nœuds immédiatement supérieurs et ainsi de suite jusqu'au nœud phrase.

Ex : Jette la petite chemise

Petit -> Adj, spatial
Petit -> Adj, jeune
Petit -> Adj, abstrait, peu important
Petit -> Adj, abstrait, méprisable
Chemise -> Nom, fém, vêtement
Chemise -> Nom, fém, classeur
Chemise -> Nom, fém, pièce de moteur

amalgame : petite+chemise ==> 3 sens sur les 9
on remonte au SN: la+petite+chemise ==> 3 sens
Jette -> V, trans, action, lancer [sujet=humain, objet=obj physique]
Jette -> V, trans, action, se débarrasser [sujet=humain, objet=obj physique]
on remonte à la P: jette+la+petite+chemise ==> 6 sens car tous les attributs sont compatibles

Critiques :

1. Weinreich : on ne tient pas compte de l'ordre des mots dans l'amalgame (professeur femme≠femme professeur et plus grave dans des phrases entières). Chomsky distingue alors le concept de structure profonde lié aux règles transformationnelles qui ne modifient pas le sens de la structure originelle et introduit des *degrés de grammaticalité*.

2. McCawley : dérivation et arbres : l'ordre d'application des règles impose la structure à l'arbre (ceci est incompatible avec l'apprentissage); les règles transformationnelles ne sont pas économiques car elles reproduisent des informations déjà contenues dans le lexique ; les restrictions de sélection : le système de traits est infini ou incomplet (ex. 'diagonaliser' ne s'applique qu'avec 'matrice')

Pour conclure on peut dire que l'amalgame-projection permet de réduire l'ambiguïté sémantique d'une phrase mais ne permet pas de trouver le sens de la phrase.

Analyse par structure sous-jacente
se fait en paraphrasant une suite d'items lexicaux à l'aide de règles transformationnelles.

Ex: x a tué y
tué->faire(aux+part. passé)+devenir+non+vivant
mais il y a un problème dans: Jean a mis Paul dehors, car mettre+dehors -> expulser -> faire(aux+part. passé)+devenir+non+odieux

Controverses : Postal puis Lakoff
problèmes du focus (accentuation des mots sémantiques), des présuppositions, du topique (thème), de l'interaction entre quantificateurs et négation. Cela conduit à une modification de la théorie standard : extension de l'interprétation sémantique aux structures de surface.

La sémantique interprétative : grammaire de cas de Fillmore

Postulat : la structure profonde n'est pas à un niveau *pertinent*. Il faut un niveau de représentation sémantique propre. Les pbs d'accord et de placement sont liés à la structure de surface (comme sujet, objet).

Le verbe a un rôle prédicatif, une phrase (ou une proposition) pouvant se mettre sous la forme Prédicat(arguments) :

je conduis mon père à Niort => conduire (agent=je, bénéf=mon père, lieu=Niort)

L'analyse se fait en envisageant les cas suivants : Agent, Expérienceur, Instrument, Action, Objet, But, Source, Bénéficiaire, Lieu, Temps, Manière --attachés aux SN de la phrase. Celle-ci s'analyse donc par :

V[cas(i), i=1 à N] où les 'cas' pointent sur les SN

on peut donc répondre par l'explicitation de cette structure aux questions : qui fait quoi, à qui, au bénéfice de qui, avec quoi, où, quand, comment, pourquoi ?

Ex : *Jean a cassé la vitre avec un marteau*
casser[agent=Jean, instrument=un marteau, objet=la vitre] notons que le mot 'avec' introduit souvent l'instrument.

Pour certaines langues (à déclinaison par ex.) il faut recourir aux cas profonds et aux cas de surface. Ces cas permettent de classer les verbes selon les cas qu'ils admettent. D'autre part on peut introduire des règles pour faire correspondre les cas profonds aux cas de surface, comme :

si ∃ Agent	alors	Agent=sujet
	sinon	si ∃ Instrument alors Instrument=sujet
		sinon si ∃ Objet alors objet=sujet, finsi
	finsi	
finsi		

Les controverses : (Lakoff) cette analyse ne fonctionne qu'en vérification.
Une phrase telle que 'Angèle a utilisé un litre de lait/pendant 2 heures/ pour faire la pâte' est correcte au sens de Fillmore. Cela montre qu'il faut revenir aux structures syntaxiques pour associer les syntagmes V, SN1 et SN2.

La sémantique logique

P(=phrase) est une fonction propositionnelle s'appuyant sur des prédicats de verbe $V(x_1,x_2,...x_n)$ où les x_i sont des arguments du SN (adjectifs, noms munis d'opérateurs logiques).
Exemple :

Paul joue un air à Dominique
P(Jouer(Paul, air, Dominique)) = vrai

Problèmes liés à la vériconditionalité :

“Ma brosse à dent est enceinte” est inacceptable (prédicat faux) tandis que "J'ai révé que ma brosse à dent est enceinte" devient acceptable car le prédicat réver(je,x) est vrai pour tout x.

“Le monsieur a embrassé la dame”
Embrasser y(x,z) ET Passé(y) ET Monsieur(x) ET Dame(z)
“Je nie que le monsieur a embrassé la dame”
comment représenter cette négation ? est-ce une négation du fait dans son ensemble ? ou que ce n'était pas un monsieur ? ou pas une dame ?

"Simone veut épouser un Italien"
1ère lecture: ∃x(tq: x=Italien ET Veut_épouser(Simone,x))
cet Italien n'est pas quelqu'un de précis
2ème lecture: Veut(Simone, (∃x(tq: x=Italien ET épouser(Simone,x))) parmi les hommes que connait Simone

"L'un de vous est sûrement en train de mentir"
1ère lecture: Etre_sûr(∃x (Mentir(x) ET x ∈ "vous"))
2ème lecture: ∃x (Etre_sûr(Mentir(x) ET x ∈"vous"))

La logique mathématique est défectueuse pour toutes les phrases introduisant une modalité sur l'assertion, comme "Paul dit que..." pense que, crois que, suppose que, sait que... *Paul croit que Simone veut épouser un Italien*

McCawley propose de traiter ce cas par emboitements, d'autres linguistes en ajoutant un opérateur spécial "iota" et d'autres enfin par introduction de la logique modale (dite naturelle) par extension de la catégorie des verbes performatifs --comme ordonner (qui introduit une forme impérative), demander (forme interrogative)-- au cas de "affirmer", "dire", "croire", "penser", etc.
Paul croît que Simone veut épouser un Italien
devient Possible(P) où P est le prédicat d'une des formes ci-dessus.

La sémantique distributionnelle

Le « distributionnalisme » est un courant de pensée apparu aux États-Unis. Il se caractérise par une accentuation du contexte grammatical et syntaxique de la langue,

sans se préoccuper du sens intrinsèque des mots. Il se concentre sur l'étude de l'ordre des mots et des règles qui régissent une langue, sans s'intéresser à la dimension sémantique profonde. Le distributionnalisme est une approche empirique, car il repose sur l'observation directe d'unités linguistiques mesurables, telles que les mots, les phonèmes et les phrases.

Il rejette donc les notions de sens et de concept, qu'il juge trop abstraites. L'analyse distributionnelle empirique est donc une analyse inductive, car elle consiste à faire des observations permettant d'induire des règles décrivant le comportement syntaxique de la langue. Plusieurs ouvrages mentionnent que le problème du contexte renvoie à l'hypothèse distributionnelle de Zellig-Harris (1954). Cette dernière postule que les mots qui apparaissent dans des contextes similaires ont des propriétés linguistiques similaires. Elle a ensuite été généralisée par John Rupert Firth en 1957.

Le distributionnalisme a donné lieu à des approches linguistiques et computationnelles qui utilisent des modèles vectoriels pour représenter le sens des mots. On la retrouve dans des modèles purement statistiques qui calculent les co-occurrences entre les mots et leurs contextes à l'aide de matrices représentant les relations distributionnelles. On la retrouve également dans d'autres modèles neuronaux fondés sur l'apprentissage automatique. La sémantique distributionnelle fournit une base théorique et pratique pour représenter le sens des mots en fonction de leur contexte. Elle constitue le fondement théorique de plusieurs modèles de plongement lexical, tels que Word2Vec, GloVe et FastText.

Les approches statistiques en traitement automatique du langage naturel ont évolué avec le temps, intégrant des modèles probabilistes, des méthodes de réduction de la dimensionnalité, ainsi que des techniques d'apprentissage non supervisé afin de mieux comprendre les relations entre les mots et les documents.

1. Modèles vectoriels catégoriques : les premiers modèles de représentation des mots étaient des modèles vectoriels catégoriques. Le «one-hot encoding» fut l'une des premières représentations catégoriques de mots. Il repose sur une représentation vectorielle binaire dans laquelle l'indice correspondant au mot, prend la valeur 1, tandis que toutes les autres cases contiennent des zéros. Le one-hot encoding a ensuite été étendu à une représentation fondée sur les sacs de mots «bag-of-words». Cette méthode consiste à représenter un document ou une phrase par une matrice indiquant le nombre d'occurrences de chaque mot Harris (1954), Salton (1971). Les modèles de représentation catégorique ont ensuite évolué pour donner naissance aux modèles à base de pondérations.

2. Term Frequency-Inverse Document Frequency : Contrairement aux modèles basés sur le poids, ils ne se limitent pas au nombre d'occurrences d'un terme. Ils prennent en compte la fréquence d'apparition d'un mot par rapport à la taille du document, à l'image des modèles «Term Frequency TF» et «Term Frequency-Inverse Document Frequency (TF-IDF)». Le TF-IDF a été introduit en 1970 comme une amélioration de la représentation des textes de la méthode «bag-of-words BoW». Il permet de pondérer l'importance des mots dans un document en tenant

compte de leur fréquence et de leur rareté dans le corpus, ce qui améliore la précision dans des tâches comme la recherche d'information Jones (1972). La fréquence du terme (TF) mesure la fréquence d'apparition d'un mot dans un document donné. C'est un indicateur de l'importance d'un mot dans un document spécifique. L'inverse de la fréquence du document (IDF) est une mesure qui réduit l'importance des mots fréquents dans tout le corpus, car ils sont généralement peu informatifs. L'idée est qu'un mot courant dans tous les documents n'apporte pas d'information spécifique Jones (1972). Le score TF-IDF pour un mot dans un document est le produit de TF et d'IDF :

TF-IDF(t, d) = TF(t, d) × IDF(t) (1.1)

Où : TF (Term Frequency) :

— TF(t, d) représente la fréquence d'apparition du terme t dans le document d.

TF(t, d) = Nombre de fois que le terme t apparaît dans le document d
Nombre total de termes dans le document d

IDF (Inverse Document Frequency) :

— IDF(t) mesure l'importance du terme t dans l'ensemble du corpus.

IDF(t) = log(Nombre total de documents dans le corpus
Nombre de documents contenant le terme t)

TF-IDF :

— TF-IDF(t, d) combine les deux mesures pour donner une mesure de l'importance du terme t dans le document d.

Bien entendu, les modèles à base de pondérations sont construits sur le modèle des sacs de mots « bag- of-words ». Par conséquent, ils partagent la même limitation : l'incapacité à capturer l'ordre des mots dans un document. Néanmoins, ils offrent de bonnes performances au niveau lexical Salton et al. (1975).

La sémantique cognitive

La sémantique cognitive considère le sens comme le produit d'opérations mentales et de structures conceptuelles partagées, plutôt que comme une simple correspondance entre mots et objets. Elle met l'accent sur les **schémas**, les **métaphores** et les **prototypes** pour expliquer comment les locuteurs comprennent et produisent le langage. Cette perspective a influencé à la fois la recherche linguistique théorique et les technologies de traitement du langage naturel, tout en suscitant des débats sur la formalisation et la prise en compte des différences culturelles.

L'une des distinctions du cognitivisme est que le sens d'un mot dépend non seulement du contexte, mais aussi des connaissances, des expériences et des représentations que les individus lui attribuent.

Jerry Fodor, l'un des piliers de la théorie de la « modularité de l'esprit », considère la cognition comme étant une représentation symbolique permettant le traitement des concepts abstraits. Dans son ouvrage intitulé « The Language of Thought », il introduit un langage symbolique mental structuré qu'il considère comme la forme de la pensée humaine Fodor (1975). Roger Schank, un autre pilier du modèle cognitif symbolique appliqué à la compréhension de la langue, soutient quant à lui que la compréhension

humaine repose sur la manipulation de structures symboliques représentant le savoir et l'expérience humaine. Ses travaux portent sur la manière dont les êtres humains interprètent le monde à l'aide de symboles. Cette conception symbolique de la pensée humaine a toutefois été critiquée par plusieurs cogniticiens, notamment Marvin Minsky, qui souligne que l'on parle d'une intelligence artificielle, néanmoins incapable d'apprentissage perceptif, d'organisation de la mémoire ou encore de raisonnement critique humain

Cartes cognitives de graphes conceptuels

D'autres courants se sont développés autour des cartes cognitives. Une carte cognitive contient deux types d'informations : des nœuds appelés états représentant des concepts et des arcs entre ces nœuds représentant des liens d'influence positifs ou négatifs. Un mécanisme d'inférence propage les influences.

Une première faiblesse des cartes cognitives est sa trop grande souplesse car un état peut être représenté par n'importe quelle étiquette linguistique. Une seconde faiblesse du modèle est l'absence de structuration des états, qui fait que des liens entre états, autres que ceux d'influence, ne peuvent pas être exprimés.

Le modèle des graphes conceptuels (Sowa 1984) est un modèle de représentation graphique de connaissances. Un graphe conceptuel est défini sur une structure appelée support permettant de spécifier en hiérarchie le vocabulaire. Une opération d'inférence, appelée projection, permet de rechercher des graphes qui sont sémantiquement liés entre eux. L'idée du modèle des cartes cognitives de graphes conceptuels consiste à décrire chaque état par un graphe. D'abord, l'utilisation d'un graphe conceptuel, associé à chaque état, permet de définir chaque état en référence à une ontologie qui est le support. Ensuite, on peut calculer ou regrouper des classes d'états qui sont liés entre eux dans une collection. Enfin, cette classification peut se combiner avec le calcul d'influence.

Un graphe conceptuel G est formé d'un ensemble de sommets concepts (C_G), un ensemble de sommets relations (R_G), un ensemble d'arêtes (E_G) et une application qui associe à tout sommet et à toute arête une étiquette ($étiq_G$). Le graphe conceptuel de la figure représente un accident mortel (accident dans lequel une personne est morte).

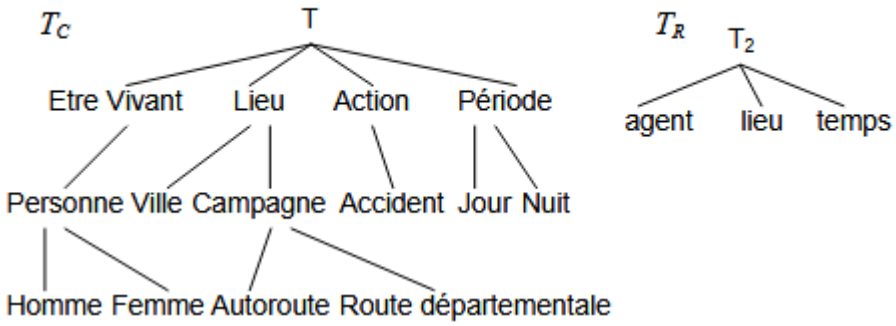


FIG. 1 – *Un support*

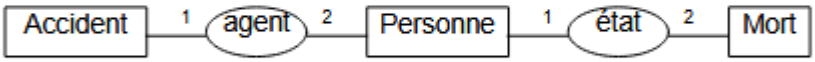
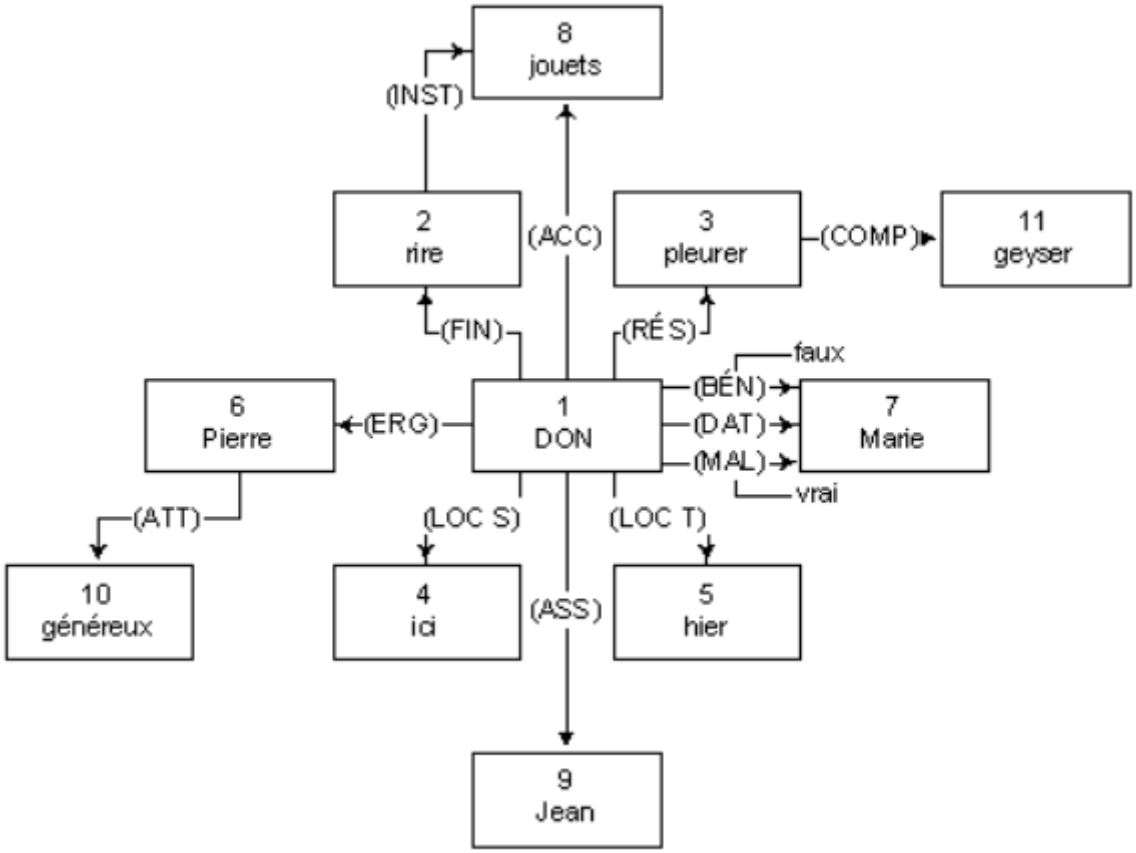


FIG. 2 – *Un graphe conceptuel*

A l'aide de ces représentations et d'algorithmes de parcours de graphes il est possible de faire des inférences pour répondre à des requêtes, par exemple ici : « un agent a-t-il constaté l'accident ? ».
Ci-après un exemple plus complet donné par François Rastier. Tout d'abord une liste de cas liés aux relations,

	CAS	DÉFINITION	DÉNOMINATION
(ACC)	accusatif	patient d'une action, entité qui est affectée par l'action	PATient
(ASS)	assomptif	point de vue	SELon
(ATT)	attributif	propriété, caractéristique	CARactéristique
(BÉN)	bénéfactif	au bénéfice de qui ou de quoi l'action est faite	BÉNéficiaire
(CLAS)	classitif	élément d'une classe d'éléments	CLASsitif
(COMP)	comparatif	éléments unis par une comparaison métaphorique	COMParaison
(DAT)	datif	destinataire, entité qui reçoit une transmission	DEStinataire
(ERG)	ergatif	agent d'un procès, d'une action	AGEnt
(FIN)	final	but (résultat, effet recherché)	BUT
(INST)	instrumental	moyen employé	MOYen
(LOCS)	locatif spatial	position dans le temps représenté (LOCS)	ESPace
(LOCT)	locatif temporel	position dans le temps représenté (LOCT)	TEMps
(MAL)	maléfactif	au détriment de qui ou de quoi l'action est faite	MALéficiaire
(PART)	partitif	partie d'un tout	PARTitif
(RÉS)	résultatif	résultat, effet, conséquence	EFFet (ou CAUse)

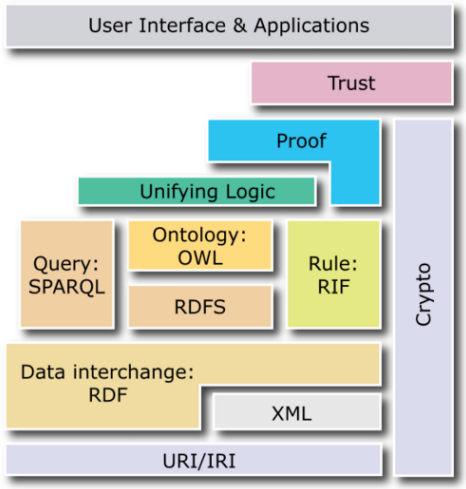
qui, appliqué à l'exemple : « Selon Jean, hier, ici même, Pierre, généreux, donna à Marie une poupée et un bilboquet pour qu'elle rit grâce à ces présents ; mais elle pleura plutôt comme un geyser. » donne le schéma :



Le web sémantique

Le Web sémantique est apparu sur la scène grâce aux recherches menées par Berners-Lee et al. (2001). Il s'agit d'une description formelle des concepts et des relations. À l'aide de standards tels que RDF, RDFS et OWL, le Web sémantique modélise des connaissances et établit des relations explicites entre concepts, enrichissant ainsi les ressources utilisées dans le TALN, telles que les ontologies lexicales et les bases de données sémantiques. Le standard *Resource Description Framework RDF* est un modèle conceptuel de données basé sur des triplets, destiné à décrire les ressources Web et leurs métadonnées Klyne et Carroll (2004). Le standard *Resource Description Framework Schema RDFS* est une extension du RDF introduisant la notion de classe et de hiérarchie entre les classes *Ontology Web Language OWL* incarne un paradigme révolutionnaire dans l'ingénierie des connaissances. Différentes définitions ont été proposées pour les ontologies : une définition des termes et des relations de base constituant le vocabulaire d'un domaine, ainsi que des règles permettant d'étendre ce vocabulaire Neches et al. (1991). Tom Gruber, quant à lui, définit une ontologie comme une description explicite de concepts et de relations, destinée à un agent ou à une communauté d'agents Gruber (1993). Pour Sowa, l'intérêt des ontologies réside dans l'étude des catégories d'objets qui existent ou peuvent exister dans un certain domaine Sowa (1995). Le Web sémantique a ouvert la porte à Plusieurs outils de traitement automatique du langage naturel ont

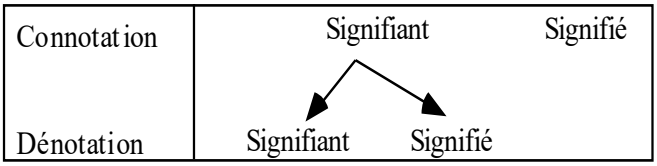
émergé grâce au Web sémantique, notamment WordNet, une ontologie lexicale contenant une base structurée de synonymes, d'antonymes, d'hyponymes et de relations sémantiques. Elle est utilisée dans des tâches telles que la désambiguïsation lexicale et l'extraction d'informations Miller (1995). OntoNotes, une ontologie combinant les annotations sémantiques, syntaxiques et discursives. Elle est utilisée pour le marquage sémantique et l'amélioration des systèmes TALN multilingues Hovy et al. (2006). FrameNet, une ontologie basée sur des cadres sémantiques, utile pour l'encodage du sens des phrases Fillmore et al. (2006). DBpedia, une ontologie multilingue pour l'extraction et la structuration des connaissances à partir de Wikipédia, Auer et al. (2007). En outre, l'interopérabilité des ressources sémantiques améliore les performances des systèmes multilingues et rend possible le développement d'agents conversationnels plus intelligents et contextuellement pertinents, révolutionnant ainsi l'application du TALN dans de nombreux secteurs.



- Le XML fournit une syntaxe élémentaire pour la structure du contenu dans les documents, mais il ne décrit pas la sémantique du document. XML n'est pas à l'heure actuelle une composante nécessaire des technologies du Web sémantique. Dans la plupart des cas, des syntaxes alternatives, comme Turtle, existent. Turtle est un standard de facto, car moins verbeux que XML, mais n'a pas été choisi à travers un processus de normalisation formelle.
- Le XSD est un langage de description de format de document XML permettant de définir la structure et le type de contenu d'un document XML. Cette définition permet notamment de vérifier la validité de ce document.
- Le RDF est un langage simple pour exprimer des modèles de données sous forme d'objets (« ressources ») et de leurs relations. Un modèle basé sur RDF peut être représenté à travers plusieurs syntaxes d'échanges, par exemple, RDF/XML, N3, Turtle et RDFa[28]. RDF est une norme fondamentale du Web sémantique[29],[30],[31].
- RDF Schema étend le RDF et son vocabulaire pour pouvoir structurer les propriétés et les classes au sein d'une ressource décrite en RDF.
- OWL ajoute plus de vocabulaire pour décrire les propriétés et les classes, comme avec les relations entre les classes, la cardinalité (par exemple « exactement un »), l'égalité, le typage des propriétés, les caractéristiques de propriétés (par exemple la symétrie), etc.
- SPARQL (prononcé *sparkle* en anglais : « étincelle »[32]) est un langage de requête et un protocole qui permettra de rechercher, d'ajouter, de modifier ou de supprimer des données RDF disponibles dans le Web à travers l'Internet.

La rhétorique

La rhétorique est l'art de persuader par le langage. La sémiologie a remis la rhétorique à l'ordre du jour à la reliant au concept de connotation qui associe aux éléments dénotés des signifiés supplémentaires.



La rhétorique étudie la forme des signifiants de connotation

Du point de vue de l'énonciation du discours, « les actes rhétoriques sont des discours (logos) qui mettent en scène leur énonciateur (ethos) en vue d'une action sur un auditoire (pathos) » ; « La rhétorique met en jeu deux niveaux de langage : le langage propre et le langage figuré. La figure (de rhétorique) est une opération qui fait passer d'un niveau de langage à l'autre. C'est supposer que ce qui est dit de façon « figurée » aurait pu être dit de façon plus directe, plus simple, plus neutre. » Au XXème siècle, avec les recherches structuralistes surtout, les figures de style quittent le terrain de la rhétorique pour devenir des éléments de la persuasion et de la communication. La linguistique moderne les classe majoritairement en quatre niveaux : niveau du mot (exemple : tropes), niveau du syntagme (exemple : oxymore), niveau de la proposition (exemple : inversions), niveau du texte (exemple : ironie)

Ces figures peuvent résulter d'opérations sur la surface de l'énoncé. On distingue aussi 4 types d'opérations - Adjonction, Suppression, Substitution, Echange - qui portent sur des éléments variants :

– Identité (cela donne les figures : *Répétition, Ellipse, Hyperbole, Inversion*),
– Similarité de forme ou de contenu (cela donne les figures : *Rime Comparaison, Circonlocution, Allusion, Métaphore, Hendiadyn, Homologie*)
– Différence (cela donne les figures : *Accumulation, Suspension, Métonymie, Asyndète*),
– Opposition de forme ou de contenu (cela donne les figures : *Attelage, Antithèse, Dubitation, Réticence, Périphrase, Euphémisme, Anacoluthe, Chiasme*),
– Fausses homologies (cela donne les figures : Double sens, Paradoxe, Antanaclase, Tautologie, Prétérition, Calembour, Antiphrase, Antimétabole, Antilogie).

Mais la rhétorique ne se réduit pas à la sémantique de phrase ou d'énoncé, elle est plutôt une manière d'organiser un discours et les liens entre parties de discours sont appelées relations rhétoriques. Initiée par Jean-Claude Anscombre et Oswald Ducrot, l'approche dite argumentative, s'efforce de restituer les actes de langage dans le contexte énonciatif. Le discours est ainsi un ensemble de présupposés et d'implicites. Néanmoins son objet reste la langue et non spécifiquement le discours, au sein desquels le locuteur comme

personne sensible et intentionnelle a une place prépondérante.

Le rapprochement entre la linguistique et la rhétorique emprunte donc deux voies : d'un côté, on envisage une rhétoricité générale en tant que condition même de l'existence de la production discursive ; de l'autre, on conçoit la rhétorique comme un instrument d'analyse discursive à la lumière des théories en cours en science du langage. Etudier l'argumentation consiste alors à observer les techniques discursives visant à provoquer ou accroître l'adhésion. Dans la filiation de la nouvelle rhétorique, l'AD (Analyse de Discours) étend la notion d'argumentation au fonctionnement discursif global : énoncer c'est argumenter. Tout discours est argumentatif dans le sens où un locuteur tend toujours à modifier la vision du monde de l'allocutaire. C'est pourquoi l'AD distingue la « dimension argumentative de la langue » de sa « visée argumentative », c'est-à-dire distingue la tendance de tout discours à influencer l'interlocuteur, à agir sur lui, de celle plus spécifique, consistant à déployer des stratégies de persuasion.

Cela amène Ducrot et coll. A mettre au point une théorie, la TBS (Théorie des Blocs Sémantiques). La TBS donne un format identique à tous les éléments constitutifs du sens. Ce qui fait sens, pour la TBS, ce sont des enchaînements de deux phrases au moyen de certains connecteurs, enchaînements auxquels est donné le nom d'« argumentations », en détournant ce mot de son sens habituel. Le schéma général de l'enchaînement argumentatif, c'est-à-dire de l'atome sémantique, est ainsi une suite X-CONN-Y, où X et Y sont des phrases et CONN est un connecteur comme *donc*, *pourtant, parce que, puisque, mais, cependant*, etc. Un bloc sémantique est alors constitué d'un carré argumentatif dans lequel on envisage toutes les possibilités logiques (négation, réciprocité, contraposition, transposition) que peut prendre cet atome sémantique selon les connecteurs envisagés.

La **Discourse Representation Theory (DRT)**, ou Théorie de la Représentation du Discours, est un cadre formel développé principalement par Hans Kamp et Uwe Reyle à partir des années 1980. Elle vise à représenter et analyser la signification des discours (textes ou dialogues) au-delà de la simple phrase, en tenant compte de la dynamique de l'information et des références entre les éléments du discours.
Points clés de la DRT

- Représentation des discours : La DRT utilise des structures appelées Discourse Representation Structures (DRS) pour modéliser le sens d'un discours. Ces structures intègrent les entités mentionnées, les relations entre elles, et les conditions imposées par le texte.
- Traitement des références : Elle permet de résoudre les anaphores (comme les pronoms « il », « elle », « cela ») en reliant les éléments du discours entre eux.
- Dynamique de l'information : La DRT considère que la signification d'un discours se construit progressivement, phrase après phrase, en mettant à jour la représentation globale.
- Logique et formalisation : La DRT s'appuie sur des outils logiques pour formaliser la sémantique des discours, ce qui la rend utile en linguistique computationnelle et en intelligence artificielle.

Exemple simple : « *Un homme entre dans un bar. Il commande une bière.* » La DRT construit une DRS où « un homme » est introduit comme une entité, et « il » est relié à cette entité dans la phrase suivante.

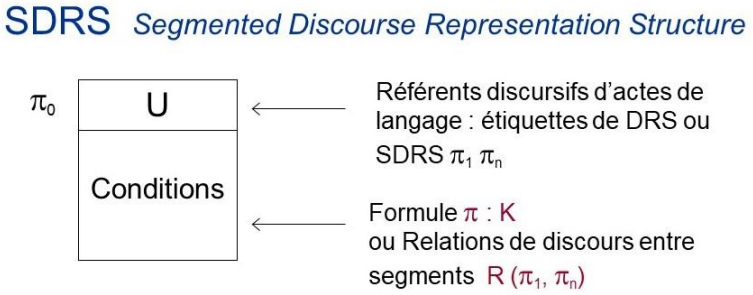
Applications

- Linguistique computationnelle : modélisation du sens pour les systèmes de traitement automatique des langues.
- Intelligence artificielle : compréhension de texte, dialogue homme-machine.
- Philosophie du langage : analyse de la signification et de la référence.

En résumé, la DRT est une théorie puissante pour comprendre comment le sens se construit et s'articule dans un discours, en combinant logique, linguistique et pragmatique.

Un prolongement de cette théorie, la SDRT (*Segmented Discourse Representation Theory)* a été proposée par Nicholas Asher et Alex Lascarides pour les discours. Les sémantiques dynamiques sont basées sur l'idée que l'interprétation du discours est un processus incrémental : l'interprétation de chaque phrase met à jour le contexte courant qui devient le contexte d'interprétation de l'énoncé suivant. Elle fournit un outil formel très riche pour modéliser la cohérence du discours en construisant une structure logique complète : la SDRS. C'est une structure hiérarchique formée de constituants reliés par des relations de discours appelées relations rhétoriques.

Une SDRS est un couple <U, Cond > où U est un ensemble de référents discursifs d'actes de langage (étiquettes p de SDRS) et Cond est un ensemble conditions sur les éléments de U.



Cette définition a pour conséquence la cohabitation de constituants simples et de constituants complexes. Comme dans la RST les relations rhétoriques lient les blocs de discours par exemple : les questions-réponses (QAP et IQAP), les questions subordonnées, les élaborations de connaissances, les délégations d'action, les élaborations de plan (Elab-q), les élaborations de question, les élaborations de but, les incidences, les répliques, etc.

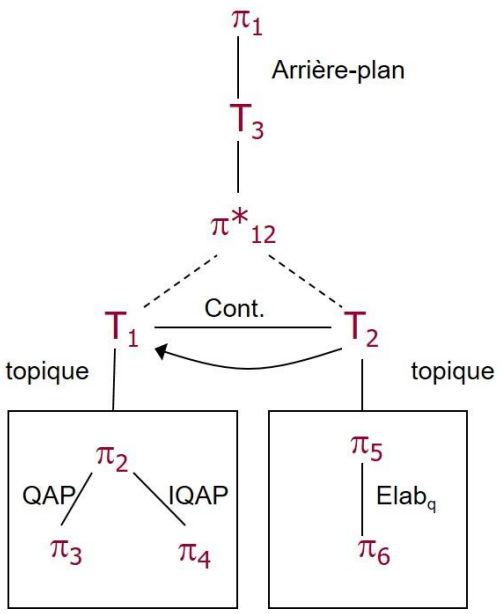
Pour étendre la SDRT au dialogue nous (Anne Xuereb et Jean Caelen) définissons le topique comme l'ensemble des éléments en cours de discussion. Sa caractérisation automatique est un problème complexe. Dans un premier temps nous considérons que le topique est caractérisé à la fois par le thème du prédicat et les champs sémantiques des référents de l'énoncé. Ces données sont codées dans une ontologie de concepts, qui

contient également la notion de compatibilité de topiques. Cette notion de compatibilité permet de détecter l'élaboration de topique, à l'œuvre dans une relation, ainsi que la clôture et le changement de topique au cours du dialogue. Nous construisons un nœud topique au-dessus de tout échange question/réponse. Ainsi le nœud topique est une SDRS marquée T dominant un échange, qui structure le fonds commun d'information établi par les participants au dialogue : il contient le résultat de la résolution de la séquence question/réponse sous-jacente, et c'est un site d'attachement disponible pour un autre segment discursif.

Exemple d'analyse pour le dialogue suivant :

A: Bonjour, Luc Blanc à l'appareil. π_1
Est-ce que la salle Apollinaire est disponible la semaine prochaine ? π_2

B : Elle est disponible jeudi et vendredi. π_3
A : Bon et bien réservez-**la moi** π_4
B : Quel jour ? Jeudi ou vendredi ? π_5
A : disons vendredi. π_6



La pragmatique

Les principaux acteurs :

Pragmatique non ling.	Pragmatique linguistique
Peirce (1839-1914)	Saussure (1857-1913)
	Russell
	Frege
Wittgenstein (1889-1951)	Bloomfield
Morris (1901)	Chomsky
Austin	
Searle	
	Benveniste
	Ducrot
	Culioli

La pragmatique linguistique :
Descartes postule qu'un langage parfait est caché sous le langage ordinaire (Port-Royal, Condillac) — la logique de Port-Royal : " les mots sont des sons distincts et articulés, dont les hommes ont fait des signes pour marquer ce qui se passe dans leur esprit " et " certaines caractéristiques du langage sont l'aboutissement de structures cognitives profondes " [Chomsky]

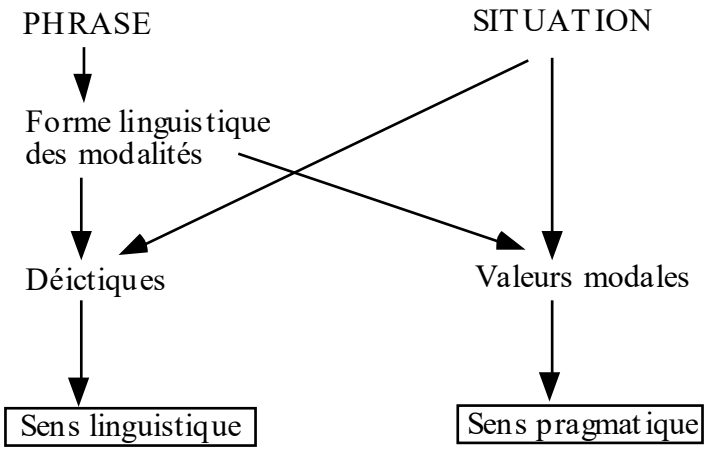
La pragmatique non linguistique :
Le signe "Vénus" synonyme de "l'étoile du berger" lui est-il substituable dans la phrase "Vénus est un mot de cinq lettres" ? Non "Vénus" a une valeur référentielle.

La pragmatique étudie l'utilisation du langage dans le discours et les marques spécifiques qui, dans la langue, attestent sa vocation discursive [Morris]. Le sens renvoie non au contenu mais à l'usage.

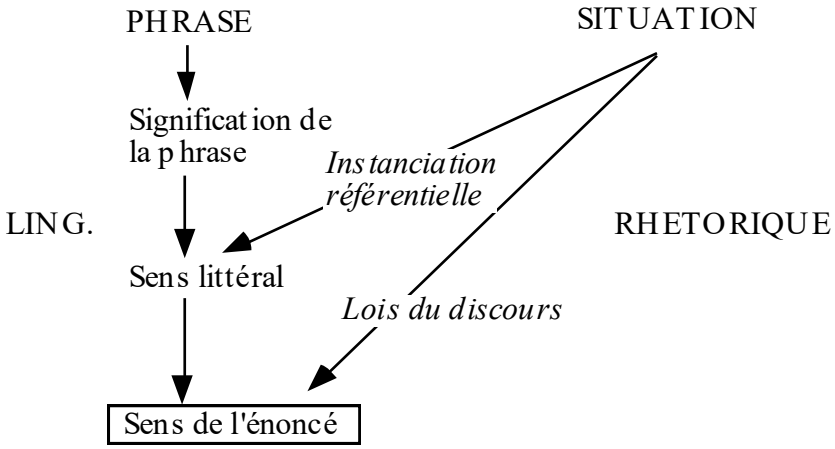
" Le concept de la pensée, du calcul, de la déduction est déterminé par un accord réalisé non pas sur les données de l'expérience incontestables (empirisme) ni sur les données d'une sorte d'ultra-expérience (platonisme) ni sur de simples définitions (conventionalisme) mais sur des formes d'action et de vie ", " parler un langage est une partie d'une activité ou d'une forme de vie " [Wittgenstein]

Pragmatique = {langue, locuteurs, monde, situation}

— Pour Morris pragmatique et linguistique interférent par les modalités énonciatives (ordre, question, assertion, etc.) et les déictiques selon le schéma suivant dans lequel les deux processus se déroulent en parallèle :



— Pour Ducrot la pragmatique s'intègre davantage à la linguistique par un double processus : celui de la signification et celui du sens (du signifiant au référent et inversement) à travers deux composantes, linguistique et rhétorique



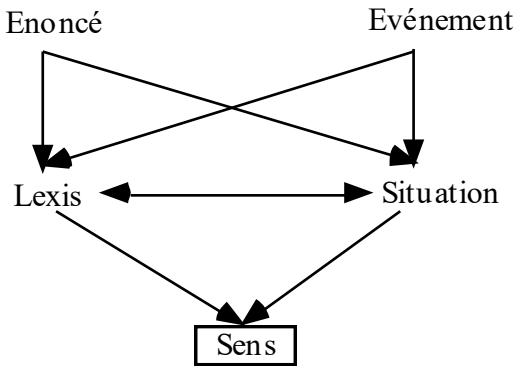
(a) Instanciation référentielle : elle s'appuie sur la distinction entre fonction référentielle et indication existentielle — "mon voisin de palier" qui entraine un présupposé sur son existence et sur le fait que j'habite dans un appartement. Ce présupposé est un élément du contenu et non une simple condition d'emploi.

(b) l'organisation interne de l'énoncé peut en modifier le sens : sous-entendu, ironie, etc. C'est à ce niveau qu'interviennent les lois du discours et les effets dus à la situation, comme modifieurs secondaires du sens.

— Searle conteste la réalité du *sens littéral :* on ne peut interpréter une phrase en dehors de tout contexte "le chat est sur le paillasson", peut-on ne pas imaginer la scène en lisant cette phrase ? (on sait par exemple que le chat ne flotte pas au-dessus du paillasson). On

doit donc préciser : sens littéral relatif à un système de présomption d'arrière-plan. Donc pour Searle la situation entre avec le sens littéral dans le processus de signification.

— Pour Culioli la référence contextuelle se fait dès le niveau lexical : il apparait que le système linguistique s'articule avec l'extra-linguistique par tout un réseau de relations entremêlées les unes aux autres, qui mettent en jeu les diverses composantes du système ainsi que l'activité symbolique et langagière de l'homme.



La référence

Comment les locuteurs désignent des entités dans le discours et maintiennent la cohérence référentielle.

La référence co-textuelle

- Jean a su la vérité par le facteur. Il lui a dit que sa femme le trompe
- Jean s'est plaint au facteur. Il lui a dit que sa femme le trompe
- Jean a plaint le facteur. Il lui a dit que sa femme le trompe.

L'attribution des référents est nettement plus complexe, elle dépend de la sémantique et de la situation.

- Cas 1 : du fait que le facteur parle à Jean, « il » réfère au facteur et « lui » à Jean, et « sa femme » est celle de Jean
- Cas 2 : ici c'est Jean qui parle au facteur, « il » réfère à Jean et « lui » au facteur, et « sa femme » est toujours celle de Jean
- Cas 3 : Jean parle au facteur, « il » réfère à Jean et « lui » au facteur, mais ici « sa femme » réfère au facteur

Ces différences d'attribution sont liées au sens des expressions « savoir la vérité par quelqu'un », « se plaindre à quelqu'un », « plaindre quelqu'un »

La référence contextuelle : la déixis

- Jean a lu ce livre puis il me l'a prêté
- Jean est arrivé à l'instant
- Jean est ici

L'attribution des référents ne peut pas se faire sans connaître le contexte et les circonstants de l'énonciation : espace, temps, personnes, objets de la scène, etc.

- Cas 1 : le déictique « ce » renvoie à l'objet « livre » par un autre canal de désignation que la langue, par exemple un geste de monstration et « me » désigne l'énonciateur
- Cas 2 : « à l'instant » renvoie au temps de l'énonciation
- Cas 3 : « ici » renvoie au lieu de l'énonciation

Ces différences d'attribution sont liées aux circonstants des expressions déictiques

Il est possible de dégager trois conceptions majeures de la deixis. Pour certains cette opération permet de rapporter les objets et les événements du monde aux coordonnées associées au locuteur : espace et temps (je-ici-maintenant = ego-hic-nunc). Pour d'autres la deixis précède ce positionnement et constitue un certain type de construction référentielle, ce qui fait de la déixis une opération énonciative qui se conjugue et s'enchevêtre avec celle que la tradition médiévale regroupe sous le terme de *modus*. Mais cette modalité langagière est-elle nécessairement associée au geste d'ostension ? Se fonde-t-elle sur la co-présence d'un champ visible où l'on trouverait sinon l'objet désigné, du moins les instructions nécessaires à la construction de son identité ? Pour d'autres enfin, la deixis est foncteur de cohésion textuelle permettant à l'orateur d'infléchir le fil de son discours en proposant un nouveau lien à l'intérêt de son auditoire. Elle se conçoit alors comme une rupture construite dans et par le discours.

1. la deixis comme repérage d'une référence constituée,
2. comme construction d'une référence,
3. comme définition d'un nouveau topos discursif.

Entre deixis et anaphore, il n'y a nullement opposition mais continuité graduée : l'anaphore est *endo*-phorique, la deixis est *exo*-phorique (phorique = qui renvoie à).
Quelques marqueurs de la deixis spatiale [Borillo, 92] : la cible est l'objet à localise, le site est le repère par rapport auquel s'établit la localisation. La forme la plus courante en français est N_{cible} V $Prép_{loc}$ N_{site} où V est un verbe statif (être, se trouver, etc.). La localisation de la cible peut être établie de deux manières :

– soit directement à partir du site (orientation intrinsèque),

> *loin de, près de, dans le voisinage, à proximité, aux alentours, dans les parages, aux abords, à l'écart, ici, là, là-bas, ailleurs, près d'ici ce côté-ci*, etc. dénotent la distance,

– soit à partir d'un autre polarisateur que le site (orientation contextuelle ou extrinsèque).
Au niveau des traits sémantiques on distingue généralement trois types de déictiques (orientés par la distance à l'énonciateur)

ICI	LA	LA-BAS
+proche	-proche	-proche
-éloigné	-éloigné	+éloigné
+locuteur	-locuteur	-locuteur
-interlocuteur	+interlocuteur	-interlocuteur

La deixis temporelle [D. Jouve, 92] : l'interprétation de "maintenant" avec le passé simple ne s'effectue pas dans un texte de fiction comme dans les énoncés de réalité; elle est étroitement liée à des phénomènes d'anaphore, car si le texte crée ses propres repères, il doit sans cesse les expliciter au moyen d'adverbes et de compléments adverbiaux qui se répondent les uns les autres en recréant sans cesse de nouveaux repères selon les points de vue privilégiés des personnages, ceux-ci pouvant bien sûr appartenir à d'autres registres que la représentation de l'humain : *Le prédicateur marque une pause; sa voix s'éleva maintenant plus lente*…

[Morel, 92] : les fonctions essentielles de la deixis (dénomination et ostension) s'organisent à partir de la fonction référentielle, qui est son attache dans le langage.

Les implicatures conversationnelles

- On a fêté l'anniversaire de Jean. Pierre aussi est venu
- On a fêté l'anniversaire de Jean. Même Pierre est venu

On peut faire des hypothèses crédibles à partir de ces énoncés

- Cas 1 : il y avait d'autres invités que Pierre et Pierre s'est rajouté à ces invités
- Cas 2 : Pierre n'était pas vraiment attendu mais il est venu

Ces différences d'attribution sont liées aux modulateurs « aussi », « même »

Les inférences plus générales :

Intentionnelles

- On a fêté l'anniversaire de Jean. Pierre est venu malgré la distance
- On a fêté l'anniversaire de Jean. Pierre est venu par hasard

On peut faire des hypothèses crédibles à partir de ces énoncés

- Cas 1 : Pierre redoute se déplacer bien qu'il ait eu l'intention de venir
- Cas 2 : Pierre n'avait pas l'intention de venir

Ces différences d'attribution sont liées aux modulateurs « malgré », « par hasard »

Circonstancielles

- Tu t'es garé où ? Juste derrière
- Tu prends la première à droite

On peut faire des hypothèses crédibles à partir de ces énoncés

- Cas 1 : garé ? Probablement un véhicule, pas un vélo, juste derrière, une rue, un immeuble, peut-être même avec de la chance
- Cas 2 : situation de conduite, copilotage, avec ordre de tourner dans la prochaine rue à droite

Ces inférences sont possibles par l'usage commun du langage et de situations habituelles fortement connotées

Non-dits, insinuations, allusions

Taire une partie de discours ou l'orienter vers un sens ou un autre

- Tu t'es garé où ? En enfer
- Tu prends toujours la première à droite…

On peut faire des hypothèses crédibles à partir de ces énoncés

- Cas 1 : détournement de la question pour indiquer la difficulté à se garer

- Cas 2 : orientation vers le côté obsessionnel du conducteur

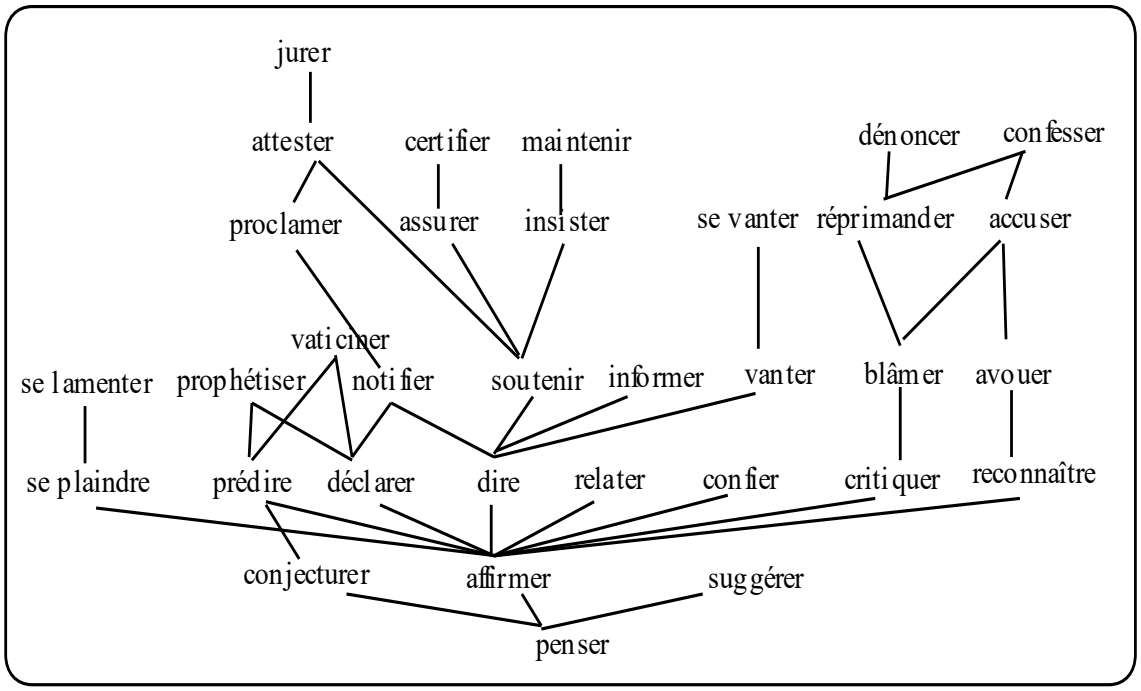
Ces inférences sont possibles par la connaissance subjective de l'interlocuteur

Les actes de langage

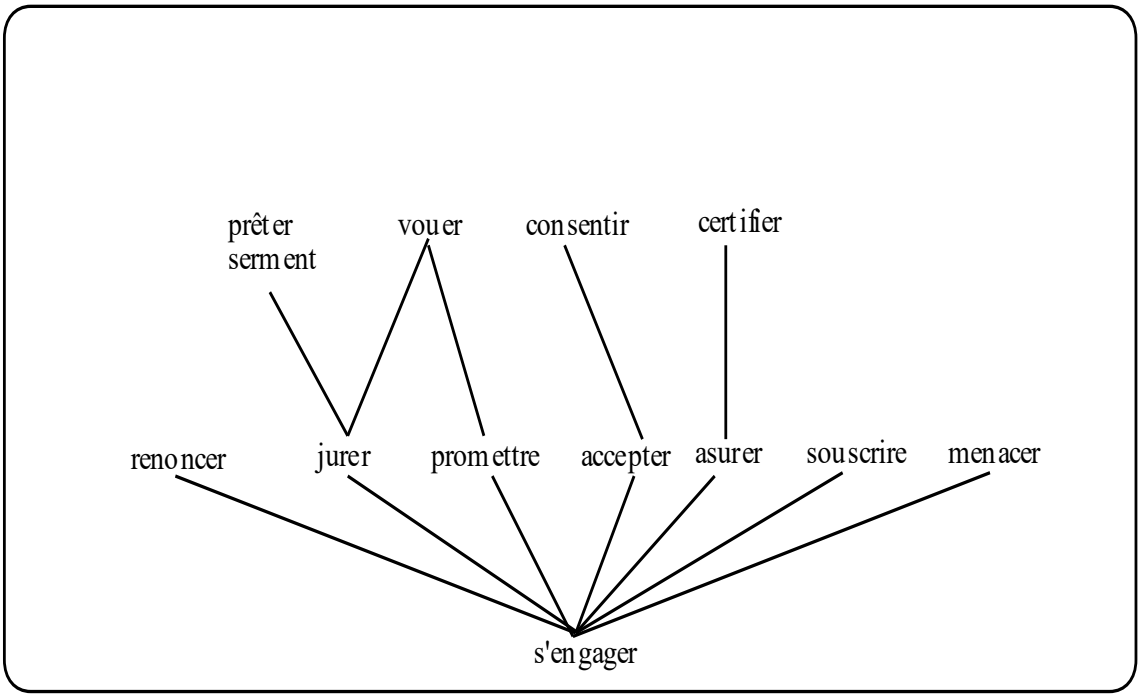
Austin puis Searle et Vanderveken définissent trois composantes dans un acte de langage : le locutoire, l'illocutoire (ou force illocutoire) et le perlocutoire munies de deux paramètres, *valeur* et *force*. Ils distinguent cinq catégories d'actes illocutoires qui pour le français sont en partie dénotées par le verbe.

1. **Acte assertif :** la composante illocutoire décrit un état de fait existant. Le locuteur dit comment sont les choses. Le but est de rendre le contenu propositionnel (qui est une proposition) conforme au monde. L'acte assertif révèle les croyances du locuteur. Nous notons cet acte FS (faire savoir).
2. **Acte directif :** le but illocutoire est de mettre l'interlocuteur (qui est ici le locuteur lui-même) dans l'obligation de réaliser une action future. Le locuteur essaie de faire faire les choses. Le but est de rendre le monde conforme au contenu propositionnel (qui contient l'action future de l'interlocuteur). L'acte directif exprime les désirs et la volonté du locuteur. Nous notons cet acte FF (faire faire une action) ou FD (faire devoir) lorsque l'obligation est forte ou FFS (faire faire savoir) pour une demande d'information.
3. **Acte promissif :** il s'agit d'une obligation contractée par le locuteur lui-même de réaliser une action future. Le locuteur s'engage à faire quelque chose. Le but est de rendre le monde conforme au contenu propositionnel (qui contient l'action future de l'interlocuteur). L'acte promissif révèle l'intention du locuteur. Nous notons cet acte FP (faire pouvoir).
4. **Acte expressif :** le but illocutoire de l'acte expressif est d'exprimer l'état psychologique qui lui est associé. La direction d'ajustement n'est pas de rendre le monde conforme aux mots ou vice versa. La proposition exprimée est présupposée : on se réjouit ou on déplore qu'elle soit vraie. Cet acte est très peu présent en DHM, nous le notons FSØ.
5. **Acte déclaratif :** le but illocutoire de l'acte déclaratif est de rendre effectif son contenu. Le locuteur provoque des changements effectifs dans le monde par ses déclarations. Cet acte a simultanément deux directions d'ajustement entre le langage et le monde. Il faut qu'il soit accompli dans une certaine institution extra linguistique qui confère au locuteur les pouvoirs de provoquer de nouveaux faits institutionnels par le seul accomplissement approprié d'actes de langage. Nous notons cet acte FA.

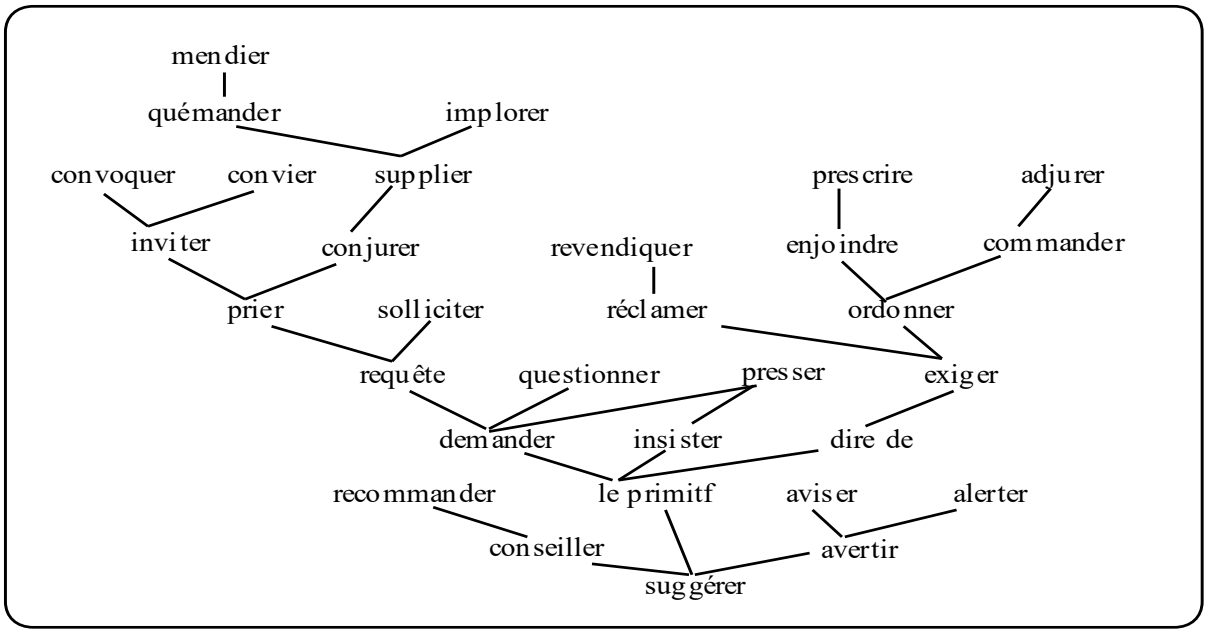
Les verbes de type assertif



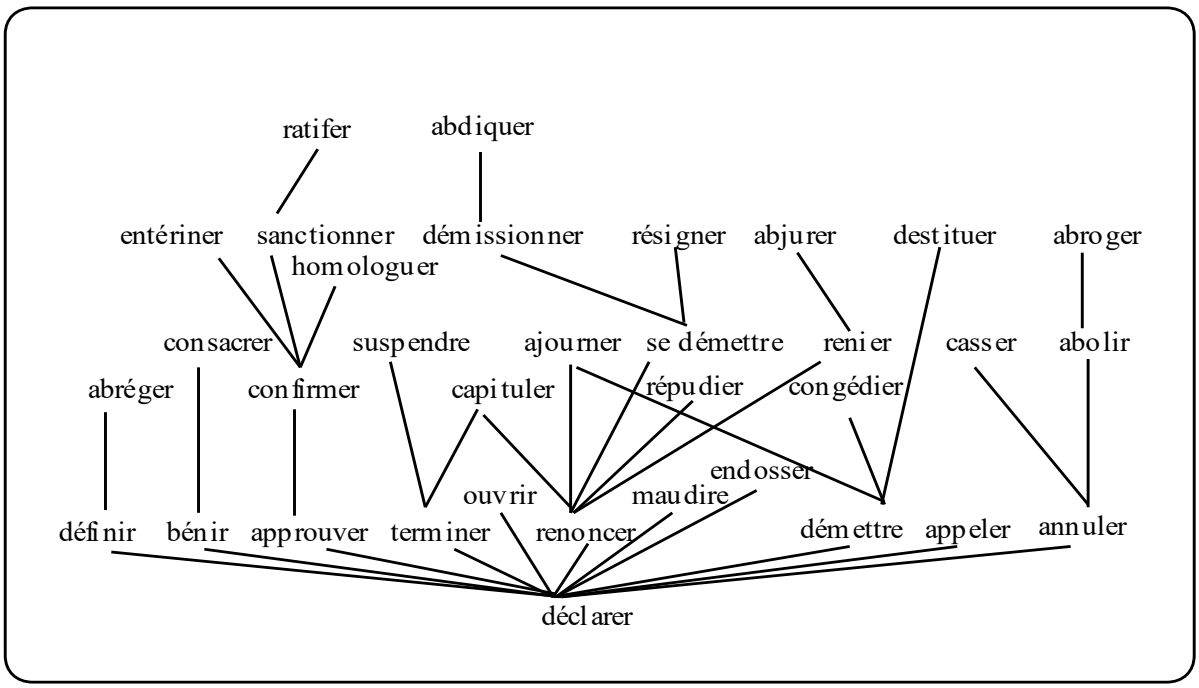
Les verbes de type promissif



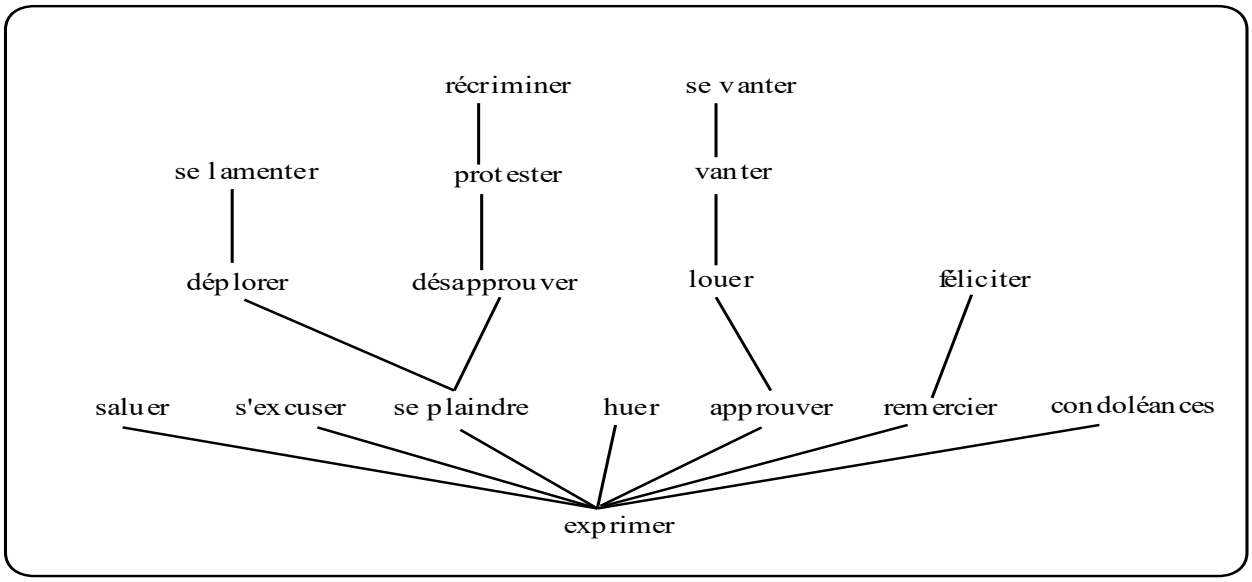
Les verbes de type directif



Les verbes de type déclaratif



Les verbes de type expressif



Ces classes sont distinctes des types d'énoncés au niveau de la grammaire classique :

1. déclaratifs : "la porte est ouverte", comment les choses sont
2. conditionnels : "il pourrait le faire" ce qui se passerait si certains faits existaient
3. impératifs : "ferme la porte", faire faire à l'allocutaire
4. interrogatifs : "comment allez-vous ?" question
5. exclamatifs : "comme c'est triste", expression des états mentaux
6. optatifs : "si seulement il pleuvait" expresion des souhaits

Pour Searle la pragmatique des actes de langage s'inscrit dans une théorie du langage et une théorie de l'action selon deux perspectives : la description des actes de langage, leur régulation selon le principe d'exprimabilité, c'est-à-dire,
(a) énoncer des mots = effectuer des actes d'énonciation, dire
(b) référer et prédiquer = effectuer des actes propositionnels, dire(p)
(c) affirmer, ordonner, promettre, etc. = effectuer des actes illocutoires, faire(p)
(d) effectuer des actes perlocutoires, faire-savoir(p')

en (c) on peut représenter tout acte de langage par F(p), p=contenu propositionnel et F=force illocutoire.
ex. : " je te promets que je viendrai ", F marquée par "je te promets" et p marqué par "je viendrai". Pour " bravo ", p=Ø.

A ce niveau Searle ne prend pas en compte le rôle du destinateur ni celui du destinataire qui n'apparaissent qu'en (d). Ceci peut lui être reproché, comme pour Chomsky : l'élimination du locuteur parlant au profit d'un locuteur abstrait. Récanati propose d'en rester au niveau (c) à des "potentialités" illocutoires qui ne sont instanciées qu'au niveau (d). En effet pour Searle l'interprétation des actes indirects pose problème. Ils se font par énonciation non littérale :

— Enonciation littérale : on dit ce qu'on veut dire « donne-moi le sel ! »,
Enonciation non littérale : on dit autre chose « peux-tu me donner le sel ? » —

Searle propose pour l'interprétation de ce type d'acte une *stratégie inférentielle,* qui examine toutes les conditions de réalisation de l'actes (situation, monde, arrière-plan, etc.), le pourquoi, les intentions du demandeur, le but poursuivi, etc., veut-il me tester dans ma capacité à passer le sel ? Cela est-il habituel de passer le sel dans ce restaurant ? veut-il me faire comprendre que je n'ai pas assez salé à la cuisson ? Etc.

La question est de savoir ici si des mécanismes plus simples fondés sur l'*usage* ne sont pas plus pertinents. Cela reviendrait à énumérer tous les cas d'usage et en fait à remplacer l'intension par l'extension, ce qui n'est pas forcément une bonne solution.

La pragmatique doit, en conclusion, relever d'une communication conçue comme "tentative d'ajustement" où l'on doit ajouter au transport de l'information, le jeu des rôles et des actes par quoi les interlocuteurs se reconnaissent comme tels, agissent comme tels et fondent ainsi des communautés linguistiques dans un monde humain. La communication est au cœur de la pragmatique.

La pragmatique est un vaste champ qui sort du TALN proprement dit. Nous n'avons fait que l'évoquer ici.

Applications

Résumé de texte

Un **résumeur de texte** est un outil qui utilise l'**intelligence artificielle (IA)** pour analyser un texte et en générer un résumé concis tout en conservant les informations essentielles. Ce processus est une application du **traitement automatique du langage naturel (NLP)** et vise à distiller les informations les plus importantes d'un document textuel plus long pour permettre aux utilisateurs d'accéder rapidement aux informations clés sans avoir à lire l'intégralité du texte

Types de résumés
Il existe deux principales catégories de résumés générés par l'IA :

1. **Résumé extractif** :
    - Sélectionne des phrases ou des paragraphes pertinents directement du texte original.
    - Utilise des techniques comme la fréquence de mots, les graphes ou les réseaux de neurones pour identifier les parties les plus importantes du texte.
    - Peut introduire des termes anaphoriques qui nécessitent de revenir au texte original pour une compréhension complète
2. **Résumé abstractif** :
    - Génère un résumé original avec des phrases qui ne se trouvent pas dans le texte d'origine.
    - Utilise des réseaux neuronaux et de grands modèles de langage (LLM) pour produire des séquences de texte valides d'un point de vue sémantique.
    - Peut paraphraser le texte original et inclure de nouveaux mots pour rendre le résumé plus concis

Processus de génération de résumés
La génération de résumés par IA suit plusieurs étapes clés :

1. **Prétraitement des données** :
    - Nettoyage du texte pour éliminer les fautes d'orthographe, la ponctuation inutile, les balises HTML, les URL, etc.
    - Tokenisation, suppression des mots vides et racinisation ou lemmatisation pour rendre le texte lisible par un modèle de machine learning
2. **Représentation des données** :
    - Segmentation et représentation des données textuelles prétraitées pour la comparaison.
    - Utilisation de modèles comme le sac de mots (bag of words), la fréquence de terme/fréquence inverse de document (TF-IDF), ou des outils de modélisation thématique comme l'analyse sémantique latente (LSA)
3. **Notation des phrases** :
    - Évaluation de l'importance de chaque phrase dans le texte.
    - Réduction de la redondance thématique pour déterminer les phrases à extraire et celles à conserver

4. **Sélection des phrases** :
    - Choix des phrases les plus pertinentes pour inclure dans le résumé.
    - Pour les résumés abstractifs, génération de nouvelles phrases qui paraphrasent le contenu original

Applications des résumés par IA
Les résumés générés par l'IA sont utilisés dans divers secteurs pour améliorer l'efficacité et la productivité :

- **Secteur du e-commerce** : Analyse et synthèse des évaluations et commentaires des clients pour comprendre les besoins et attentes des clients
- **Secteur des médias** : Résumé automatique des articles de presse pour faciliter la lecture sur des terminaux mobiles
- **Secteur financier** : Résumé des documents internes, contrats juridiques et rapports financiers pour une meilleure gestion des connaissances

Avantages des résumés par IA

- **Gain de temps** : Accès rapide aux informations clés sans avoir à lire l'intégralité du texte.
- **Efficacité opérationnelle** : Amélioration de la qualité du travail en fournissant des informations condensées et claires.
- **Personnalisation** : Possibilité de générer des résumés adaptés à des besoins spécifiques en fournissant des instructions claires à l'IA

Bonnes pratiques pour des résumés pertinents
Pour obtenir des résumés de qualité, il est recommandé de :

- **Sélectionner les bons textes** : Privilégier les documents riches en informations pour obtenir des résumés pertinents.
- **Formuler des instructions claires** : Fournir des directives précises pour guider l'IA dans la génération du résumé

En résumé, les résumeurs de texte, basés sur l'IA offrent une solution efficace pour gérer la surcharge d'informations et améliorer la productivité dans divers domaines.

Générateur de texte

Un **générateur de texte** est un outil d'**intelligence artificielle (IA)** qui utilise des **modèles linguistiques** (LLM) pour produire automatiquement des contenus textuels. Ces modèles sont entraînés sur de vastes corpus de textes afin de comprendre et reproduire les structures linguistiques, les styles d'écriture et les contextes sémantiques

Fonctionnement technique
Le processus de génération de texte repose sur plusieurs étapes clés :

1. **Entrée (Prompt)** : L'utilisateur fournit une **invite** ou un **prompt**, qui est un texte initial ou une instruction décrivant le contenu souhaité
2. **Traitement par le modèle** : Le modèle analyse le prompt et utilise des **algorithmes de traitement du langage naturel (NLP)** pour prédire les mots

ou séquences de mots suivants. Ces algorithmes sont souvent basés sur des **réseaux neuronaux profonds**, comme les architectures **Transformer**

3. **Génération de texte** : Le modèle génère une séquence de mots en prédisant, mot par mot, la suite la plus probable. Cette prédiction est basée sur les données d'entraînement et les règles linguistiques apprises
4. **Sortie** : Le texte généré est produit en fonction des instructions initiales. La qualité du résultat dépend de la précision du modèle et de la qualité du prompt

Applications

Les générateurs de texte sont utilisés dans divers domaines, notamment :

- **Création de contenu** : Articles, blogs, posts sur les réseaux sociaux
- **Marketing** : Rédaction de publicités, emails, et contenus promotionnels
- **Assistance à l'écriture** : Aide à la rédaction, correction, et amélioration de textes
- **Traduction et localisation** : Adaptation de textes dans différentes langues
- **Chatbots et assistants virtuels** : Interaction avec les utilisateurs via des dialogues automatisés

Avantages

- **Gain de temps** : Automatisation de la rédaction, permettant de produire du contenu rapidement
- **Efficacité** : Réduction des efforts nécessaires pour créer des textes de qualité
- **Personnalisation** : Adaptation du style et du ton en fonction des besoins spécifiques

Défis et limites

- **Cohérence et pertinence** : Les modèles peuvent parfois produire des textes incohérents ou hors contexte
- **Biais et éthique** : Risque de reproduire des biais présents dans les données d'entraînement
- **Hallucinations** : Génération de contenu factuellement incorrect ou inventé

En résumé, les générateurs de texte sont des outils puissants qui transforment la manière dont le contenu est créé, offrant des avantages significatifs tout en présentant des défis à relever.

Traduction automatique

Un traducteur automatique fonctionne selon les principes suivants :

1. **Analyse du texte source** :
    - Le texte à traduire est analysé pour en extraire le sens, la structure grammaticale et les mots-clés.
    - Cette étape peut inclure la segmentation du texte en phrases et en mots.
2. **Transfert** :
    - Le sens du texte source est converti en une représentation intermédiaire qui peut être comprise par le système de traduction.
    - Cette représentation est ensuite transformée en une structure linguistique cible.

3. **Génération du texte cible** :
    - Le système génère un texte dans la langue cible en respectant les règles grammaticales et syntaxiques de cette langue.
    - Cette étape peut inclure l'ajustement de l'ordre des mots et la conjugaison des verbes.
4. **Post-traitement** :
    - Le texte généré peut être soumis à des corrections pour améliorer la fluidité et la cohérence.
    - Cela peut inclure des ajustements orthographiques, grammaticaux et stylistiques.
5. **Apprentissage automatique (pour les systèmes modernes)** :
    - Les traducteurs automatiques modernes utilisent souvent des techniques d'apprentissage automatique, comme les réseaux de neurones, pour améliorer la précision des traductions.
    - Ces systèmes sont entraînés sur de grandes quantités de données bilingues pour apprendre les correspondances entre les langues.
6. **Feedback et amélioration continue** :
    - Certains systèmes permettent aux utilisateurs de fournir des retours sur la qualité des traductions, ce qui aide à améliorer les algorithmes de traduction.

Les traducteurs automatiques peuvent varier en complexité et en précision en fonction de la technologie utilisée et de la qualité des données d'entraînement.