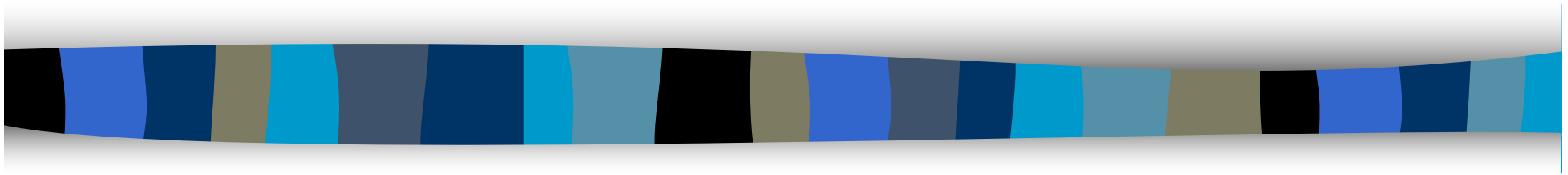
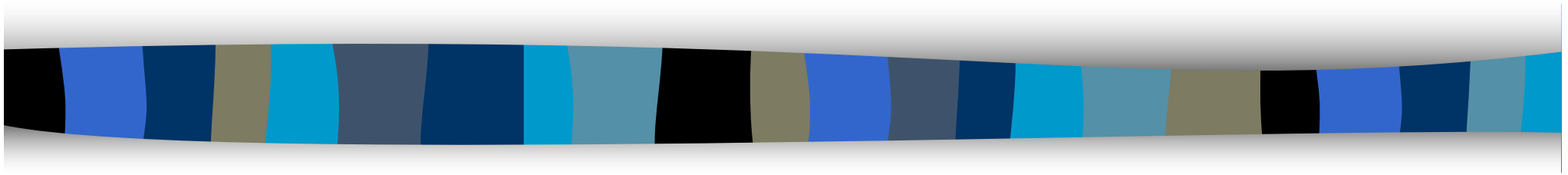
HAI815I Langage Naturel 1 (syntaxe)
Intro: modéliser le langage



Christian Retoré

Université de Montpellier
Equipe TEXTE du LIRMM

Modéliser le langage: linguistique et informatique, une longue histoire fructueuse



La linguistique computationnelle au carrefour de l'informatique, de la linguistique,

des sciences cognitives

et des outils, notamment pour Internet.

Plan

- Linguistique Computationnelle: historique
- Quelques applications
- Les modules de la linguistique
- Grammaire générative
  principes, hiérarchie, acquisition
- Logique et sémantique
  richesse des questions classiques

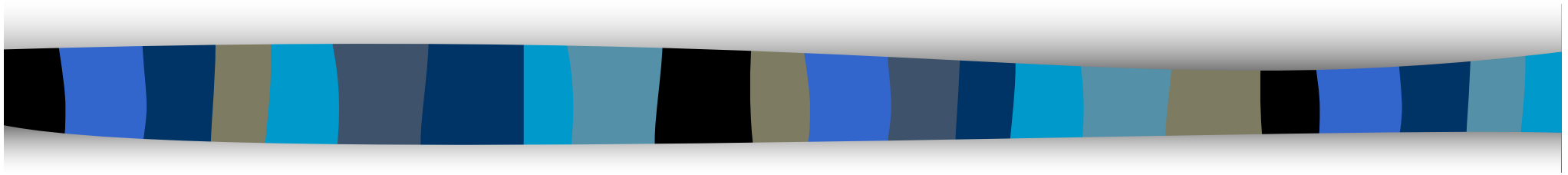
Linguistique et informatique: une longue histoire (1/2)

- 1949 Machine Translation aide à la traduction,
  - The flesh is weak but the spirit is willing
  - The meat is rotten but the vodka is strong
  - La chair est faible mais l'esprit est fort.
- 1963 aux USA arrêt ou plutôt réorientation des recherches de ce style,
  suite à un rapport de Yehoshua Bar-Hillel.

Linguistique et informatique: une longue histoire (2/2)

- Aujourd'hui réparti dans les deux disciplines suivantes
  - 1960 Computational Linguistics
    Structuration du précédent (théories mathématiques, linguistiques)
  - 1965 Automatic / Natural Language Processing
    Focalisé sur les outils et plus particulièrement:
    - Analyse syntaxique
    - Méthodes statistiques
  - 1970 Natural Language Understanding (AI)
    approches cognitives

La linguistique computationnelle



vue à travers quelques outils

Quelques outils issus de la linguistique computationnelle (1)

- Le Graal: la traduction automatique
  (il faut savoir tout traiter pour y parvenir)
- Aide à la traduction:
  - domaine spécifique
  - repère les expressions idiomatiques (aller bon train)
  - propose pour chaque mot ou expression des traductions
  - les assemble avec les choix du lecteur
  - (éviter au maximum la représentation des connaissances)

Quelques outils issus de la linguistique computationnelle (2)

- L'interface homme/machine en langue naturelle par exemple:
  - interrogation de BD en langage naturel
    Quels sont les films des années cinquante qui passent actuellement à Bordeaux?

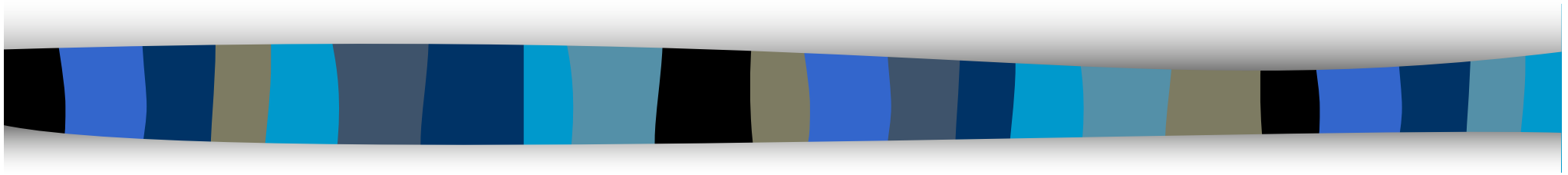
Quelques outils issus de la linguistique computationnelle (3)

- Correcteurs orthographiques (pas simple):
  - Synapse Word (souligné vert: français, italien,…)
  - Quels livres crois-tu qu'il sait que je pense que tu as lus ?
- Génération automatique de bulletins météo, de comptes-rendus,..
- Résumé automatique:
  deux techniques contrastées

Quelques outils issus de la linguistique computationnelle (4)

- Recherche d'information
  (notamment sur Internet)
  - production laitière / production de lait
  - production minière / production de mine(s) ???
- Reconnaissance de la parole
  (par ex. pour sous-titrage)
  nécessite une analyse morpho-syntaxique
  pour fonctionner en temps réel

Linguistique computationnelle



Méthodes et objectifs

Un domaine interdisciplinaire

- Mathématiques
  - Logique et théorie des langages
  - Probabilités
- Informatique
  - Algorithmique
  - Génie logiciel
- Linguistique
  - Grammaire générative
  - Descriptions linguistiques
  - Philosophie du langage

Des objectifs variés

- Réalisation d'outils de traitement des langues
- Formalisation des théories linguistiques vérification ou réfutation d'hypothèses
  - Par ex. modèles syntaxiques analysables et apprenables efficacement (en temps polynomial)
- Développement des théories informatiques et mathématiques pour elles-mêmes, éventuellement pour d'autres objectifs
  - Par ex. Théorie des langages et bioinformatique

Un aperçu des domaines de la linguistique

Diviser l'objet d'étude
en aspects plus simples

Voix et sons

- Phonétique: étude des sons concrets d'une langue
  - Accoustique
  - Système phonatoire/auditif
- Traitement du signal / médecine
- Phonologie
  - Les sons abstraits: systéme discret (dans un continu)
  - Bali / Paris indistincts pour un japonais
- Théorie des langages / automates

Prosodie (module transverse)

- Structure du phrasé et de leur enchaînement: pauses, intonation
  - "Je serai très heureux de venir parler au LaBRI, laboratoire auquel je dois ma formation initiale en informatique, par exemple sur la lambda-DRT."
  - "Je serai très heureux de venir parler au LaBRI --- laboratoire auquel je dois ma formation initiale en informatique --- par exemple sur la lambda-DRT. »
- Sytèmes d'annotation généralement superposé à d'autres informations

Morphologie: structure des mots (1/2)

- morphologie dérivationnelle: formation des mots
  - préfixes, suffixes, nom composés, etc.
  - changement de catégorie possible
    - noble ➔ noblesse
    - petit ➔ petitesse
    - maison ➔ maisonnette
    - camion ➔ camionnette
    - carpe ➔ carpette ?
- Théorie des langages / automates

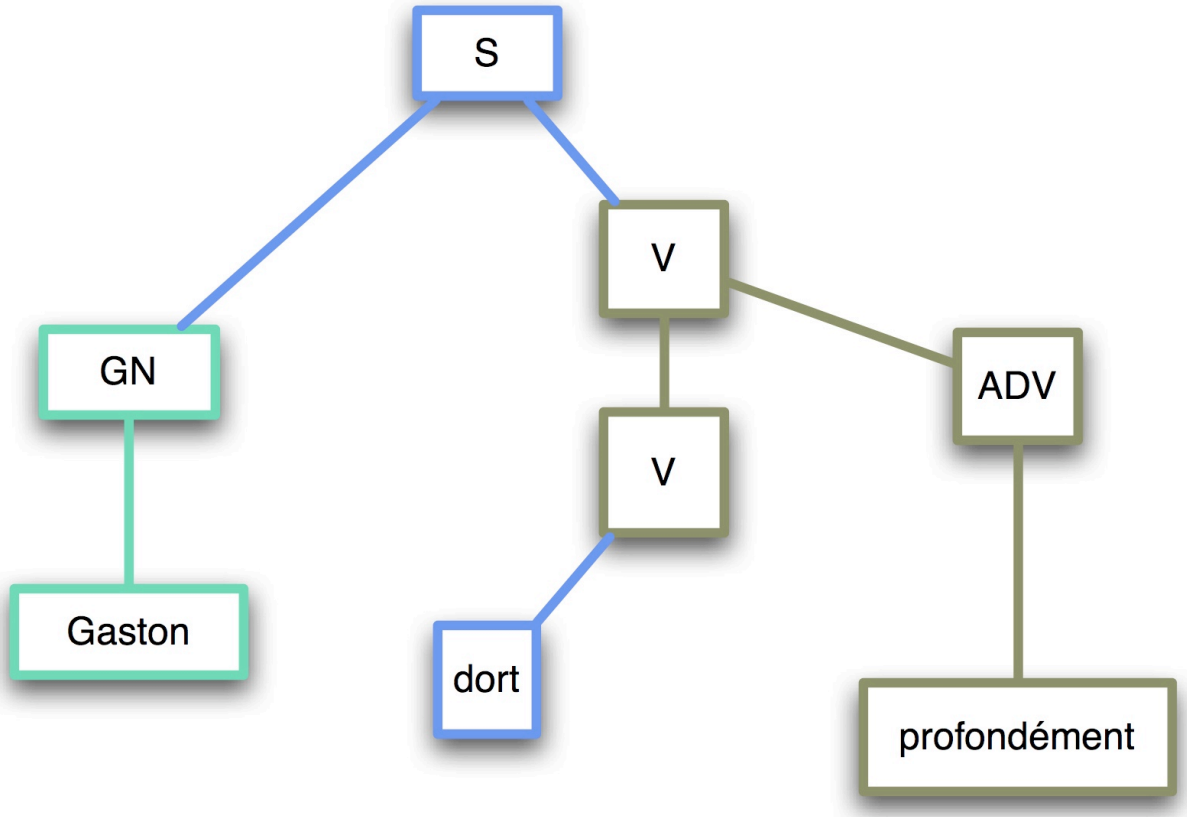
Morphologie: structure des mots (2/2)

- morphologie flexionnelle
  déclinaisons, conjugaisons
  - en général pas de changement de categorie (sauf exceptions, par exemple participe passé)
    - arriver ➔ arriv[er][ons]
    - cheval ➔ chevaux
- Théorie des langages, automates

Syntaxe

- Analyse de la structure de la phrase
  - *Je fais la réparer
  - Je la fais réparer
  - * [[Pierre [mange une]] pomme]
  - Pierre [mange [une pomme]]
- Théorie des langages de chaînes, d'arbres voire de graphes

Syntaxe



Sémantique lexicale: les sens des mots et leurs relations

- Exemple Livre:
  - livre, imprimer (objet concret),
  - lire (contenu abstrait)
  - Rôle télique: être lu, informer, cultiver,
  - Rôle constitutif: pages, convertures etc.
  - Rôle agentif: imprimeur,…
- Logique, probabilités

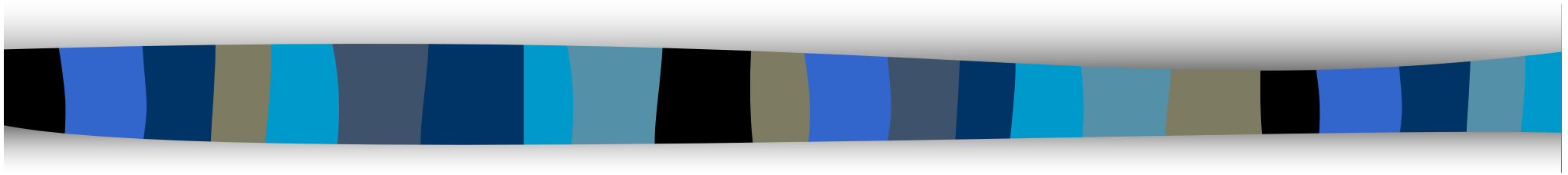
Sémantique logique (phrase, discours, dialogue, …)

- Deux aspects indépendants:
  - Sémantique Vériconditionnelle: (sens = formule logique et interprétations dans des mondes possibles) Le sens d'un énoncé c'est l'ensemble de ses conditions de vérité.
  - Une vache regarde le train passer.
  - Sémantique Compositionnelle (sens = formule logique ou construction abstraite) On calcule le sens d'un constituant, d'une phrase, d'un discours d'un dialogue à partir du sens de ses constituants et de sa structure (syntaxique, discursive,…)
  - J'ai oublié à l'hôtel ce livre que j'ai beaucoup aimé.
- Logique

Pragmatique: sens et contexte énionciatif

- Discours, dialogue, …
- Référence des indexicaux: 1ère et 2e personnes (je, nous, vous), ici, maintenant, ce, cette, . . .
  - Allons plutôt dans ce restaurant.
- En cours de formalisation, extensions des méthodes logiques en sémantique

La grammaire générative



Théorie incontournable

du XX^{e} siècle

La grammaire générative

- Théorie linguistique aussi utilisée en
  - Informatique (compilation, parallélisme)
  - Mathématiques (théorie des groupes)
  - Biologie(génomique)
- Origines
  - (Panini, Inde, V^{e} siècle avant J.C.)
  - Noam Chomsky 1955

Rupture avec le « behaviorisme »

- Une langue N'EST PAS l'ensemble des énoncés produits par les locuteurs
- MAIS
- Un ensemble fini de règles inconscientes
  qui permet de produire ces énoncés.

Rupture avec le « behaviorisme »

- Arguments:
  - Soit P la phrase la plus longue à ce jour,
    il croit que P est sans doute aussi une phrase.
  - Règles inconscientes:
    - le jeune enfant dit « vous faites »
      puis « vous faisez » puis « vous faites ».
      Le « vous faisez » ne peut provenir que de règles
      (surgénéralisation).
    - Il a aimé trois des livres qu'Echenoz a écrit.
      Il ≠ Echenoz.
    - Le chien de Paul pense qu'il ne l'aime pas.
      Il≠I tout le reste est possible.

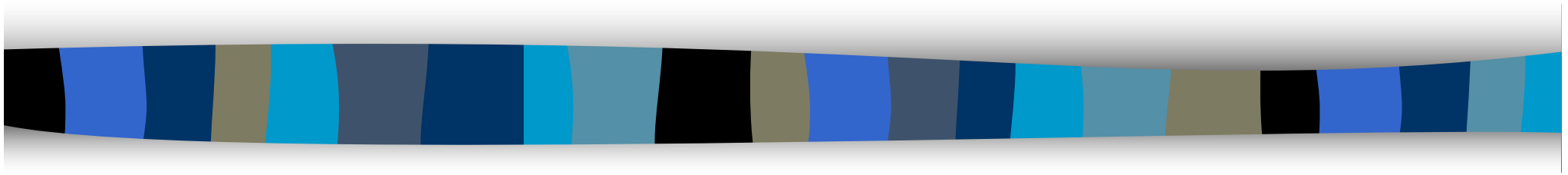
Deux principes

- Les phrases sont analysables (compréhensibles) en temps raisonnable (traduction informatique: en temps polynomial)
- Il existe un bon algorithme d'apprentissage de la grammaires à partir d'exemples positifs en nombre relativement faible.

Compétence / Performance

- Les règles / Nos limites (mémoire)
  - Le loup a dévoré la chèvre.
  - La chèvre que le loup a dévoré avait mangé le chou.
  - ? Le chou que la chèvre que le loup a dévoré avait mangé appartenait au passeur.
  - ?? Le passeur auquel le chou que la chèvre que le loup a dévoré avait mangé appartenait possède plusieurs bateaux.
  - ??? Les bateaux que le passeur auquel le chou que la chèvre que le loup a dévoré avait mangé appartenait possède sont des barges.
- (néanmoins correct, en prenant son temps et un crayon)

Quels langages formels pour la syntaxe du langage naturel



Seulement le principe de complexité de l'analyse et d'adéquation descriptive

Quelles règles modélisent la compétence?

- T terminaux (mots), N non terminaux
- Règles W → W’ (W: au moins un N)

= {
  - W=W1 Z W2 and W’= W1 W’’ W2 contextuelles
  - |W’|≥|W| croissantes
  - |W|=1 non contextuelles
  - |W|=1 et W’=mZ régulières

Exemple de grammaire hors-contexte

s ➔ sn sv

sn ➔ det n | np | det n rel _s | det n rel _o | np rel _s | n p rel _o

rel _s ➔ pro_s sv

rel _o ➔ pro_o sn vt

sv ➔ vi | vt sn

pro_o ➔ que

pro_s ➔ qui

vt ➔ regard e | regardent | mange | mangent

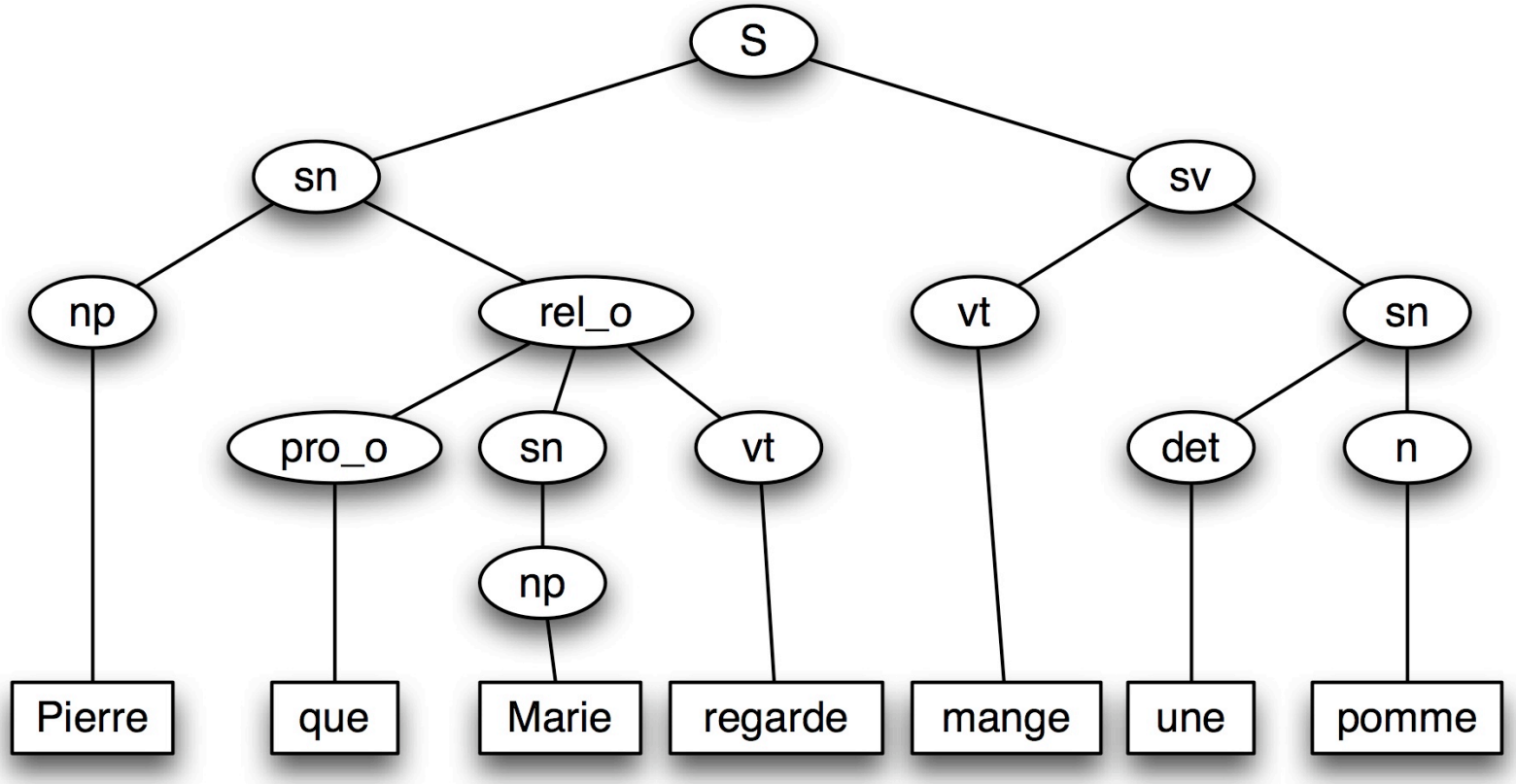
vi ➔ dort | dorment | tombe | tombent

det ➔ une | un | la | le | des | les

n ➔ pommes | pomme | femme | femmes

np ➔ pierre | marie

Exemple de dérivation



Quel type de règles?

- les langages réguliers ne suffisent pas:
  - (ex. précédent relative avec « que »)
  - Sujet1 Sujet2 Sujet3 ...
    Verbe3 Verbe2 Verbe1

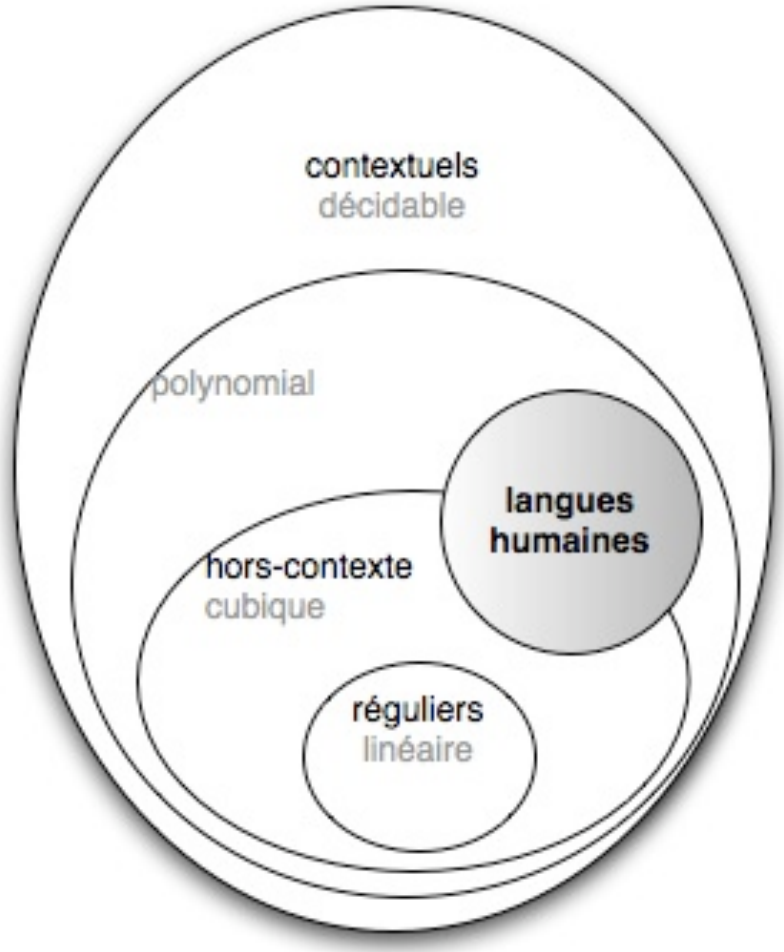
Quel type de règles?

- les langages hors-contexte non plus:
  - (complétives NL)
    Sujet1 Sujet2 Sujet3 ...
    Verbe1 Verbe2 Verbe3
  - …dat ik$_1$ Henk$_2$ *haar*$_3$ de nijlpaarden$_3$ zag$_1$ helpen$_2$ voeren$_3$
    … que je$_1$ vois$_1$ Henk$_2$ l$_3$ aider$_2$ à nourrir$_3$ les hippopotames

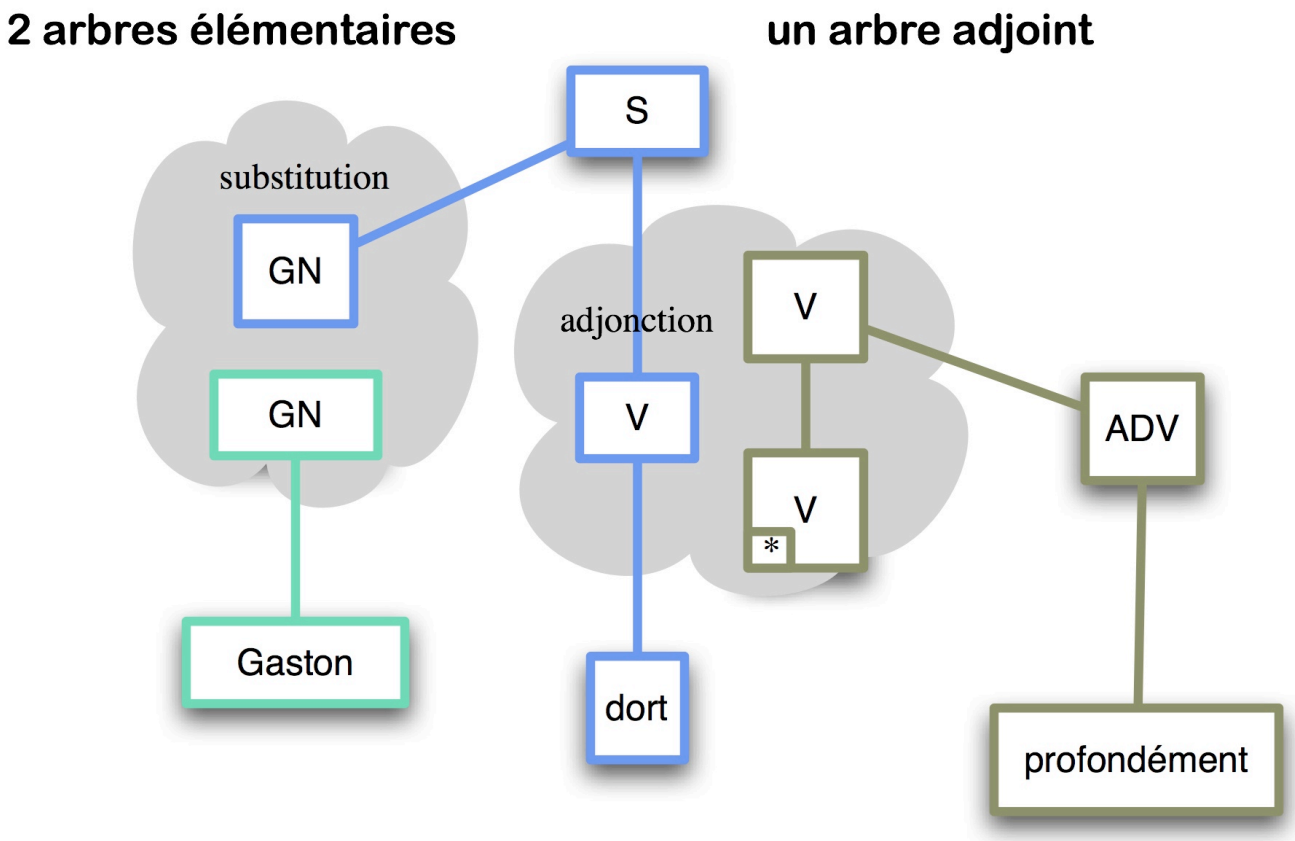
Quel type de règles?

- Un peu plus complexes que hors-contexte, mais avec analyse polynomiale :
  - TAG ou grammaires hors-contexte avec mouvements

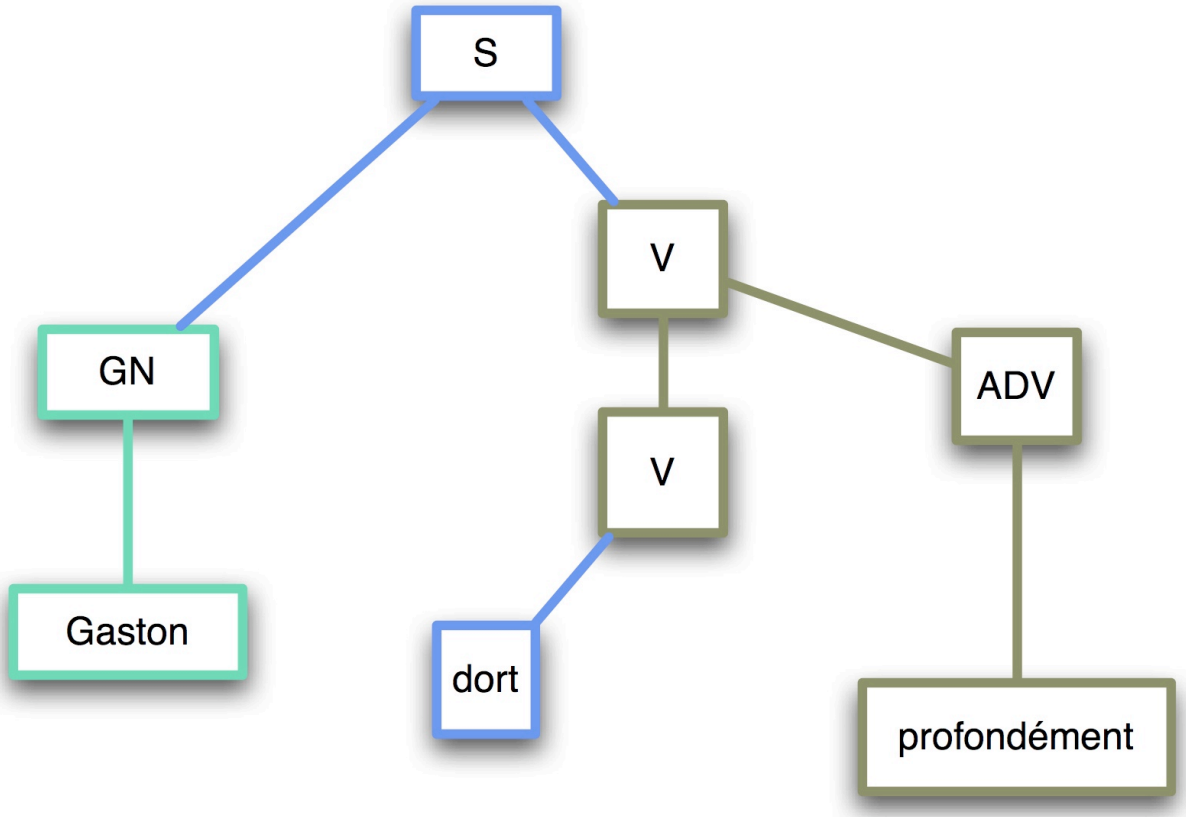
Hiérarchie des langages formels



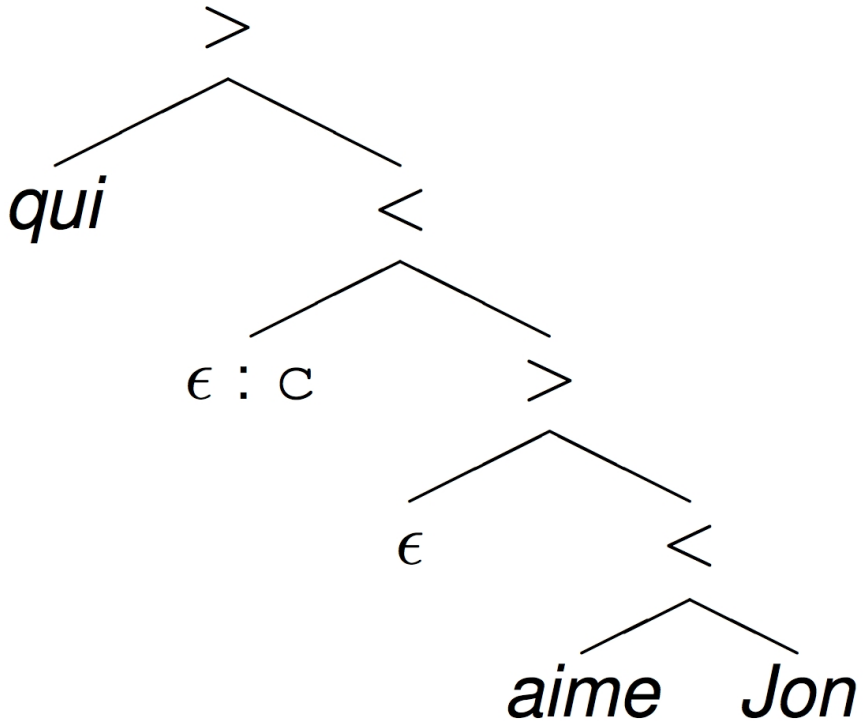
Grammaires d'arbres adjoints



Grammaires d'arbres adjoints

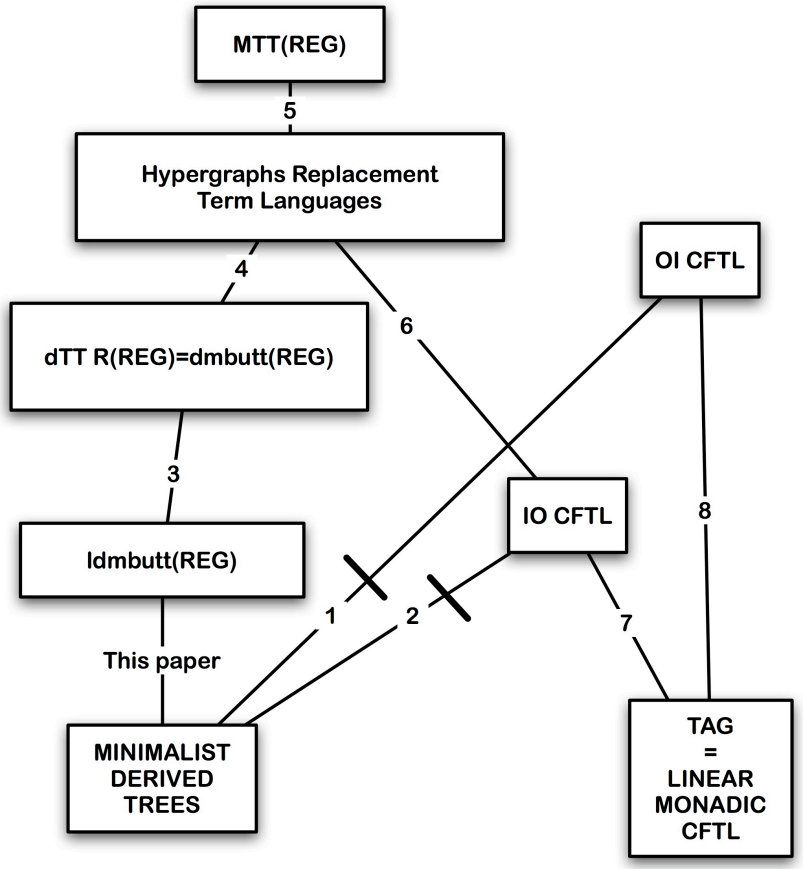


Arbres minimalistes

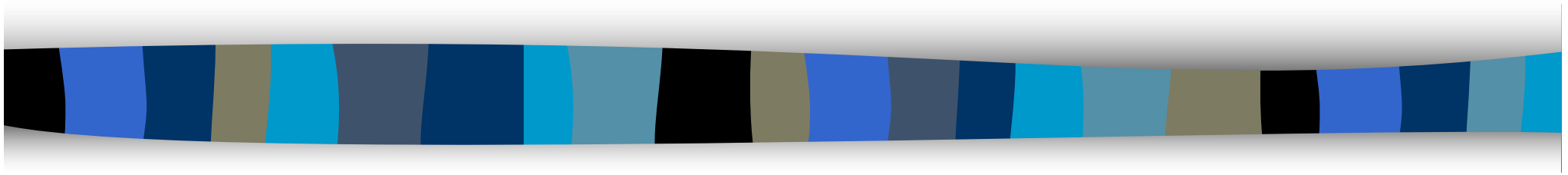


Hiérarchie des langages d'arbres

- Sujet de recherche actuel (Los-Angeles, Berlin, Bordeaux,…)



Gros plan sur un analyseur à large échelle GRAIL (Richard Moot)



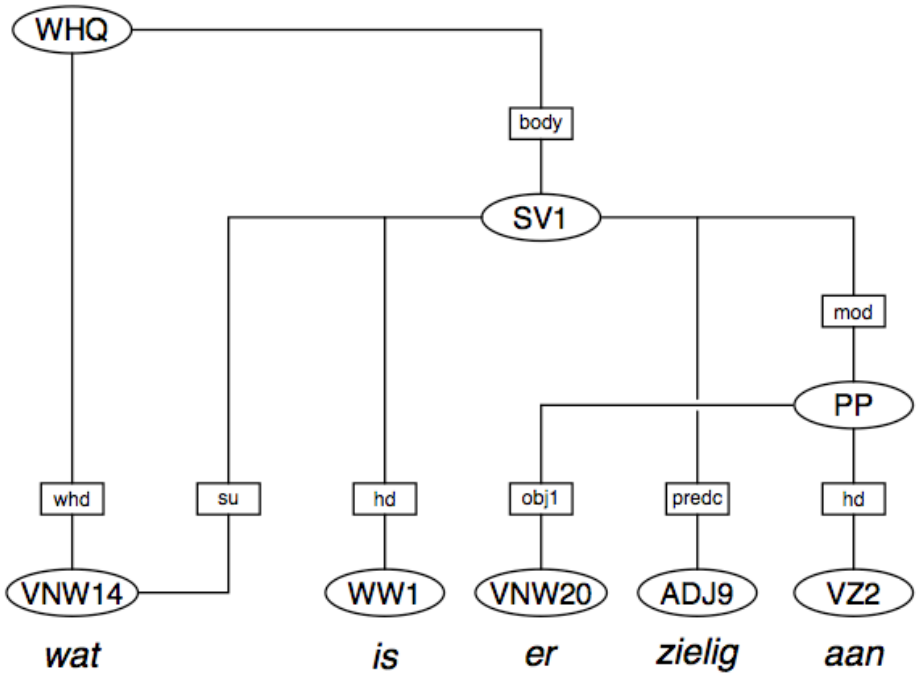
Extraction automatique de la grammaire

Association mots entrées lexicales

Analyses des séquences les plus probables

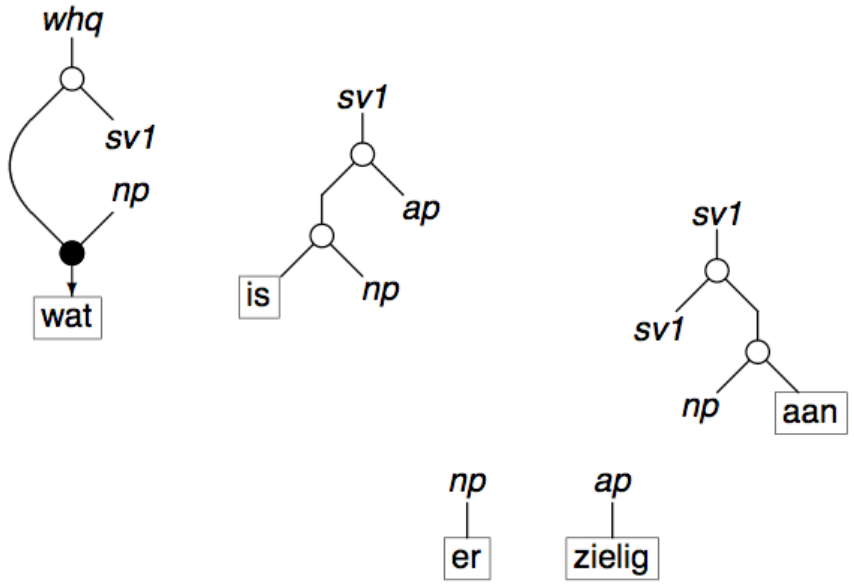
Un gros plan: Grail (Moot) 1/4

- Corpus de néerlandais parlé, transcrit et annoté (dépendances, constituants) [origine: NWO]



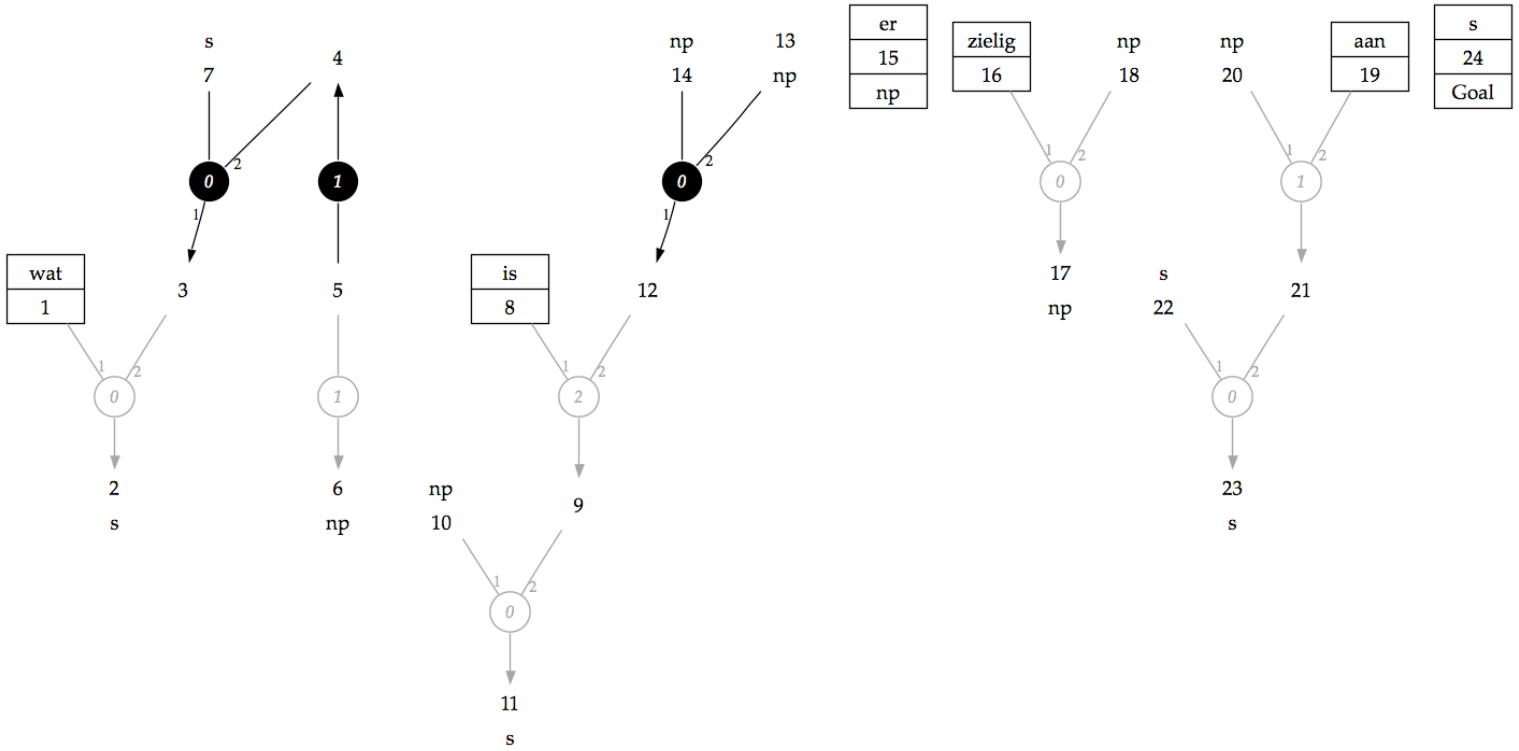
Un gros plan: Grail (Moot) 2/4

- Extraction automatique d'arbres lexicaux (techniquement: formules de la logique multimodale non associative)



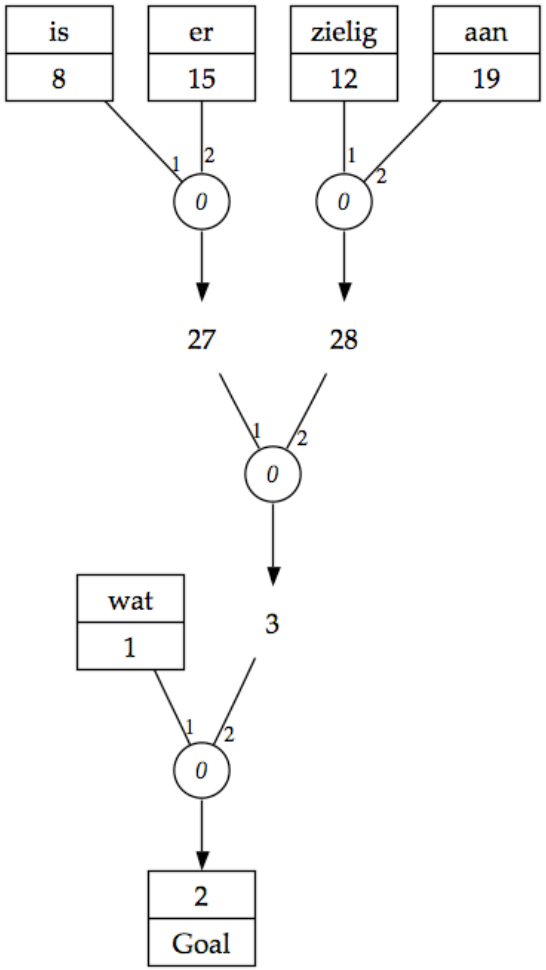
Un gros plan: Grail (Moot) 3/4

- Supertagging: étiquetage le plus probable de la suite de mots par des arbres (environ 100 arbres par mot…)



Un gros plan: Grail (Moot) 4/4

- Analyse de la phrase
  - Analyse avec chacune des n suites d'arbres les plus probables
  - Minimisation de la somme des distances des liens qui établissent la consommation des traits grammaticaux.



Etat de l'art en pratique

- Richard Moot GRAIL MMCG: extraction, parsing
  - Initialement Néerlandais NWO Dutch Spoken Corpus
  - Multi-Modal Categorial Grammar, extraite automatiquement
  - Français Corpus Paris 7 → Annotations Tigra -> catégories (en moyenne 100 arbres pas mot!)
  - Supertagging (les n plus probables suites d'assignations d'arbres aux mots de la phrase)
  - Analyse des 7 meilleures suites de supertags
    dans 96% des cas la bonne analyse est parmi les 7

Etat de l'art en pratique

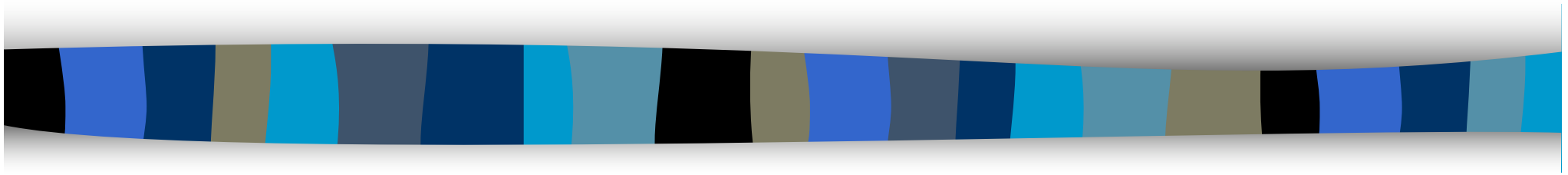
- Benoît Sagot, Eric de la Clergerie LFG parsing
  - Corpus EASy (Evaluation des Analyseurs Syntaxiques) Journaux, web, mail, discours politiques, littérature,…
    - 87177 mots
    - 4322 phrases (20,2 mots par phrase)
  - Grammaire LFG écrite "à la main"
  - Choisit une analyse par phrase
  - temps d'analyse: total 152s, 35ms/phrase 1,7ms/mot
    - Tronçons (chunks) corrects: 86%
    - Relations correctes entre tronçons (chunks): 49%
    - Analyses correctes en nombre inconnu (pas de corpus de test)

Analyseurs difficilement comparables

1. Ecrit
2. Phrases assez longues, une vingtaine de mots
3. Annotations élémentaires
4. Grammaire écrite
5. Lexical Functional Grammar
6. Mesure de correction: % tronçons et relation

1. Parlé
2. Phrases courtes mais tordues
3. Corpus bien annoté
4. Grammaire acquise automatiquement
5. MultiModal Categorial Grammar
6. Mesure de correction: % analyses complètes correctes

Acquisition de la syntaxe et grammaire universelle



Après l'efficacité de l'analyse,

Deuxième guide de la grammaire générative.

Faits connus sur le jeune enfant 18-36 mois

- Exemples positifs seulement
- Exemples ssez peu nombreux par opposition à complexité d'une langue naturelle
- Sens des mots connu au préalable
- Intonation utile

L'hypothèse de la grammaire universelle

- Pas une grammaire au sens usuel mais des contraintes sur la forme des grammaires des langues humaines
- Avec cette hypothèse le processus d'acquisition devient explicable
- Apprentissage par choix de paramètres
- Exemple bête SVO ou SOV?
  un exemple suffit:
  - maman conduit la voiture

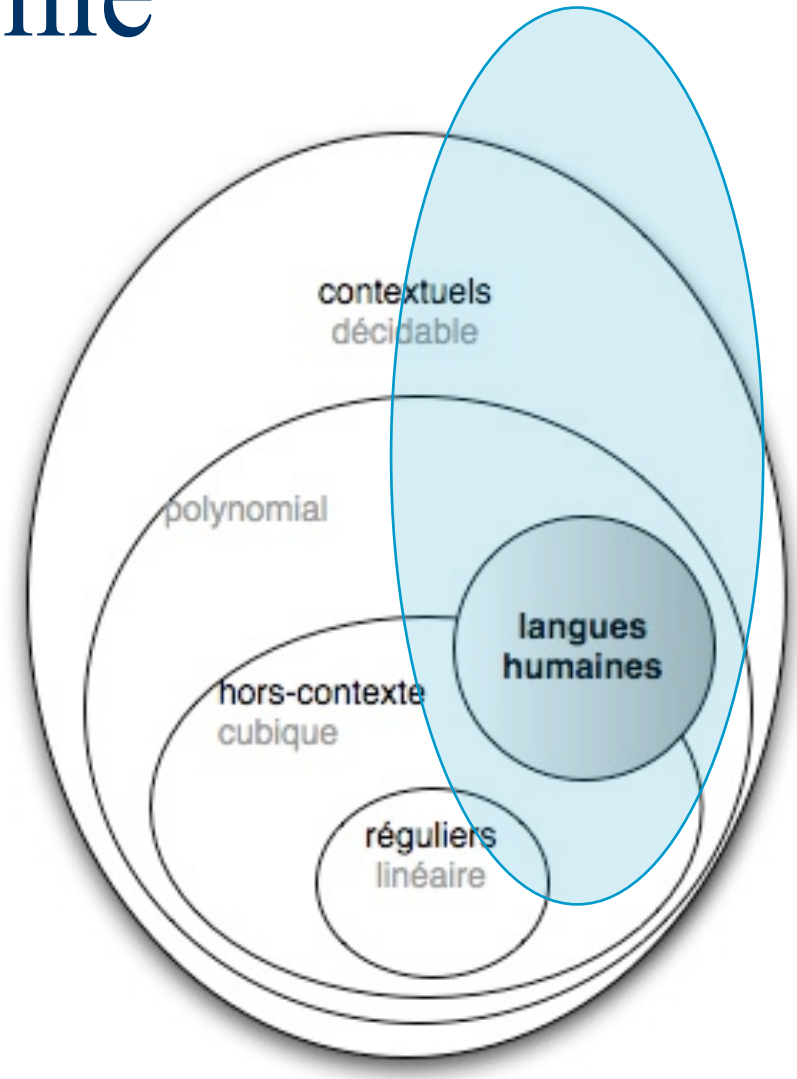
Quelques principes de la grammaire universelle

- Tout groupe nominal doit recevoir un cas,
- et seul un verbe conjugué donne un cas.
  - Il semble que l'été arrive.
  - L'été semble arriver.
  - * Il semble (que) l'été arriver.
- Un pronom doit être gouverné par son antécédent (position relative dans l'arbre d'analyse)
  - * Il_i a aimé deux livres que $Chomsky_i$ a écrit.
  - Combien de livres que $Chomsky_i$ a écrit a-t-il_i aimés?

Modèle de Gold: classe apprenable

- Algorithme
  $Phrase_1$…. $Phrase_n$ ➔ Grammaire G_n
  G_n engendre $Phrase_1$…. $Phrase_n$
- SI la totalité des $Phrase_1$……
  énumèrent un langage L de la classe
- ALORS à partir d'un nombre fini n d'exemples, l'hypothèse faite G_n ne varie plus et le langage engendré par G_n est L.

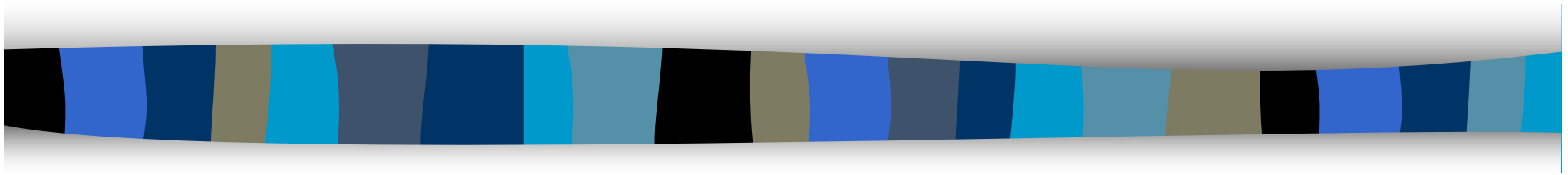
Langages apprenables et hiérarchie



Limites de l'approche à la Gold

- (Pratique) Nécessite des exemples avec beaucoup d'informations, très structurés: données disponibles?
- (Théorique) Les algorithmes, dits par généralisation, font grossir le langage jusuq'à la cible. (Sauf I. Tellier U. Lille3), à partir de représentations sémantiques: apprentissage par spécialisation)

Linguistique et logique



Une longue tradition

Aujourd'hui: l'analsye du sens devient (en partie) calculable.

Comprendre les phrases analysées

- Arbres, structures calculées automatiquement, mais ensuite, comment leur associer des formules manipulables par une machine?
- Quels problèmes rencontre-t-on?
  Ils sont très classiques
  … mais pas faciles.

Logique et grammaire: un lien naturel et traditionnel

- Depuis l'antiquité (Aristote, Denis de Thrace)
- puis au Moyen-Âge (scholastique),
- au 18e(Port-Royal)...

- La phrase a une structure logique, importante en pratique.
  - Les enfants prendront une pizza
  - Chaque en prend une pizza pour lui?
  - Ils partagent la même?

Les langues naturelles sont logiquement (trop?) riches

- Un, des certains, Tous, tous, les chaque
- + d'autres quantificateurs: la plupart, les, un grand nombre de, un petit nombre de,
  - La plupart des politiciens ont lu un livre d'économie.
- Les nombres sont aussi des sortes de quantificateurs:
  - Mettre huit gouttes dans trois cuillèrées à soupe d'eau.
  - 3x8=24 gouttes?
  - 8 gouttes?

Problèmes de portées, suite

- Lectures de re et de dicto
  - James Bond croit que l'un des chercheurs du laboratoire est un espion.
  - James Bond pense que Blofeld est un espion.
  - Il existe un espion x et JB croit que x fait partie du laboratoire.
  - James Bond a trouvé un microfilm dans le laboratoire.
  - JB croit qu'il existe un espion dans le laboratoire, mais il ne soupçonne personne en particulier.

Interprétation, mondes possibles, intentionnalité

- Sens d'un énoncé (véric onditionnel) classe des mondes possibles dans lesquels il est vrai.
  - Cet étudiant croit que Chomsky est informaticien.
  - Dans tous les mondes possibles compatibles avec les croyances de cet étudiant, Chomsky est informaticien.

Une logique d'ordre supérieur

- Propriétés de propriétés
  - Une fraise rouge vermillon
  - Rouge propriété, vermillon propriété de propriétés
  - tous les médecins sont des conducteurs
  - (donc) tous les médecins bordelais sont des conducteurs bordelais
  - *(donc) tous les bons médecins sont des bons conducteurs
  - Bordelais, médecin: propriétés
  - Bon: propriété de propriété, transformateur de propriété

Compositionnalité

- Frege le sens du tout est construit à partir du sens des parties.
  - Les étudiants reçus sont partis fêter ça.
- Limites de la compositionnalité:
  - Si un paysan possède un âne, alors il (=le paysan) le (=l'âne) bât.

Attention à ne pas s'écarter de ce qui est dit

- J'avais trois trombones dans ma poche, je les ai tous perdus sauf un.
  Je le range dans un tiroir.
- J'avais trois trombones dans ma poche, j'en ai perdu deux.
  * Je le range dans un tiroir.
- Pourtant, d'un point de vue purement logique, la situation est identique.

Lien avec la syntaxe

- Aspects sémantiques des catégories syntaxiques
- Les catégories ou parties du discours ont une contre partie logique.

Lien avec la syntaxe

- Groupes nominaux: individus (individus ou variables d'individus quantifiables) E
- Verbes, groupes verbaux: prédicats
  dort: fonction de E dans T E → T
  regarde de E dans E dans T: E → E → T
- Adjectifs partage les caractères avec les noms (accord, déclinaisons) et avec les verbes (expriment un prédicat) plutôt E → T
- Groupes prépositionnels :
  ni des prédicats, ni des individus
  - Modificateur de prédicat:
    sur un banc (E → T) → (E → T)
    dort → dort sur un banc

Grammaires catégorielles

Mot	Catégorie Syntaxique	Type Sémantique
Pierre	SN	E
Dort	SN\S	E → T

Pierre dort	S	T	Dort(Pierre)

Perspectives

- Pratiques: constitution et interopérabilité des ressources, formalismes de haut-niveau
- Théoriques et pratiques: développement des aspects sémantiques
- Théorique: que cela dit-il de nos capacités cognitives (par ex. Quelle logique utilisons nous et comment?)

Bibliographie

- The language Instinct de Steven Pinker traduit aux éditions Odile Jacob L'instinct de langage
- Cori M et Léon J, "La constitution du TAL. Etude historique des dénominations et des concepts", Traitement Automatique des Langues, n° 43-3 :21-55. 2002.
- Ch. Retoré Les mathématiques de la linguistique computationnelle.Premier volet: la théorie des langages. Deuxième volet: logique. *La Gazette des mathématiciens* N° 115 et 116 http://smf.emath.fr/Publications/Gazette/index.html

Dans ce cours
Langage Naturel 1 (syntaxe)

- Modèles de Markov Cachés ~ automates avec des probas
  Etiquetage grammatical avec l'algo de Viterbi
- Hiérarchie de Chomsky, grammaires hors-contexte (CFG), algo d'analyse syntaxique (Violaine Prince)
- Analyse syntaxique avec un réseau lexical sémantique (Mathieu Lafourcade)
- Grammaires de Clauses Définies DCG (extension des CFG) avec des traits (accord syntaxique et sémantique)
  gros TP