Analyses morphologique et syntaxique

Cours M1 DAC
UPMC

Plan du cours

- analyse morpho-syntaxique
- analyse syntaxique
- applications

Analyse morphosyntaxique

Morphologie en linguistique

- morphologie en linguistique :
  - domaine qui traite de la structure interne des mots
  - linguistique structurale :
    - notion de morphème = unité linguistique minimale (ie non décomposable) porteuse de sens
      - unités abstraites
    - notion de morphe = (une) forme graphique d'un morphème
      - allomorphes : variantes d'un même morphème
        - libres : assois/assieds
        - contextuelles : j'/je

(Berhard, 2006)

Morphologie en linguistique

- Procédés morphologiques
  - flexion : déclinaison, conjugaison
    - grand/grands/grande, cours/courir
  - dérivation : formation de nouveaux mots notamment par adjonction d'affixes au radical
    - anti-constitu-tionn-elle-ment
  - composition : combinaison de plusieurs bases pour former un nouveau mot
    - tournevis

Analyse morphologique

- Racinisation (*stemming*)
  - but : supprimer la terminaison des mots
    - congugaison/conjuguer → conjug
  - très utilisé en recherche d'information
- Lemmatisation
  - but : ramener les variantes flexionnelles d'un même mot à sa forme canonique, le lemme
    - conjugue/conjuger/conjugué → conjuguer
- Décomposition
  - but : segmenter un mot contenant plusieurs autres mots afin de retrouver ses composants
  - surtout utilisé dans des langues comme l'allemand

Analyse morphologique

- Segmentation
  - but : découper un mot en segment morphémiques
- Analyse morpho-syntaxique
  - but : analyser chaque mot pour lui associer divers types d'informations telles que la catégorie grammaticale, des traits morphologiques ainsi que le lemme correspondant

Catégories morpho-syntaxiques

- catégories de mots
  - catégories/étiquettes morpho-syntaxiques, tags, parts-of-speech...
  - cf. grammaire scolaire: noms, verbes, adjectifs, préposition...

Classes ouvertes/fermées lexicalement

Classes ouvertes

Noms

Propres

IBM
Italie

Communs

chat/chats
neige

Verbes

voir
enregistré

Adjectifs *gros petite*

Adverbes *lentement*

Nombres

122,312
un

…

Classes fermées

Déterminants *le du*

Conjonctions *et car*

Pronoms *il celui-ci*

Prépositions *de avec*

Particules *off up*

Interjections *Oh Hé*

…

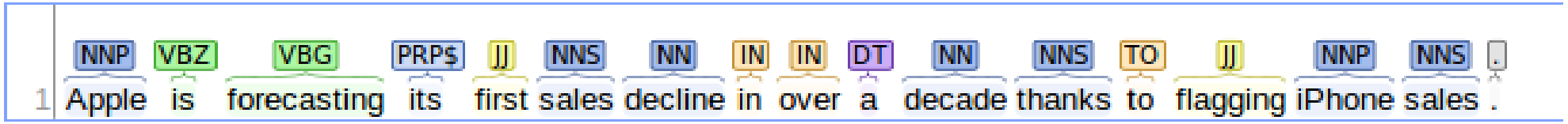
(cours C. Manning)

Etiquetage morpho-syntaxique

- les mots ont généralement plus d'une étiquette possible
  - Le bois vient de France. → le=det, bois=nom
  - Je le bois. → le = pronom, bois = verbe

- Objectif de l'étiquetage: déterminer l'étiquette pour une instance d'un mot

Part-of-Speech:



http://corenlp.run/

Exemples d'étiquetage et difficultés

- Entrée: Le débat est relancé.
  - ambiguités: le=det/pro débat=verbe/nom est=verbe/nom
- Sortie: Le/DET débat/NOM est/VER relancé/VER .

- Applications:
  - synthèse vocale: comment prononcer *est* ?
  - recherche dans un corpus: *est* en tant que nom
  - entrée d'un analyseur syntaxique
  - ...

Performance d'étiquetage

- Combien d'étiquettes sont correctes ? précision
  - étiqueteurs sur l'anglais autour de 97%
  - mais baseline simple = 90%
    - chaque mot du lexique → étiquette la plus fréquente
    - mots inconnus → noms
  - beaucoup de mots ne sont pas ambigus
    - déterminants, prépositions, ponctuation...

Déterminer l'étiquette peut être difficile pour des humains également

- Un principe décliné dans la loi relative à l'informatique
- Les statistiques ethniques, c'est complètement has been
- La Commission nationale de l'informatique et des libertés (Cnil) étudie au cas par cas les demandes

Sources d'information

- Sources d'information
  - contexte des mots
  - Le bois vient de France
    - DET NOM VER PREP NAM
    - PRO VER VER PREP NAM
  - connaissance des probabilités d'étiquettes des mots

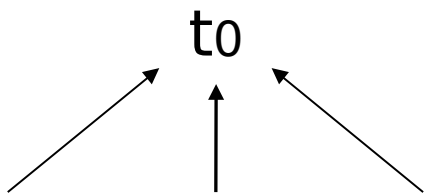
Exemples de performance de modèles

- Quelques précisions (sur l'anglais)
  - étiquette la plus fréquente: ~90%
  - trigramme HMM:
  - maxent: 94%
  - MEMM: 97%
  - dépendances bidirectionnelles: 97%
  - borne supérieure: ~98% (accord interannotateur humain)

Etiquetage avec/sans information contextuelle

Baseline

3 mots



Modèle	Caract.	Mots	Inconnus	Phrases
Baseline	56 805	**93,69%**	82,61%	26,74%
3mots	239 767	**96,57%**	86,78%	48,27%

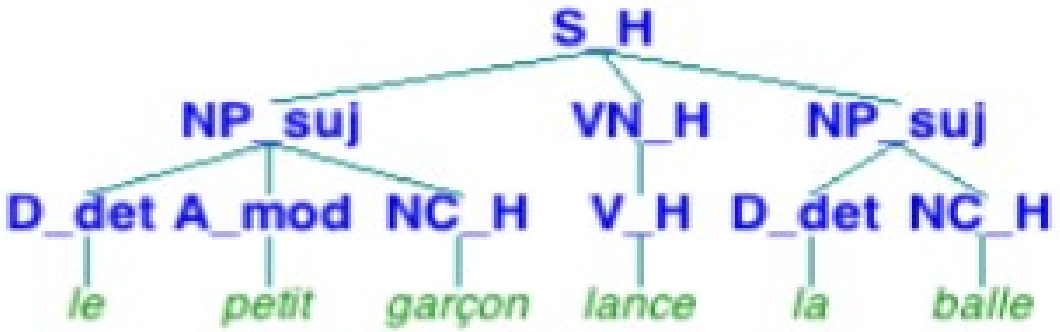
mots uniquement ~ modèle HMM

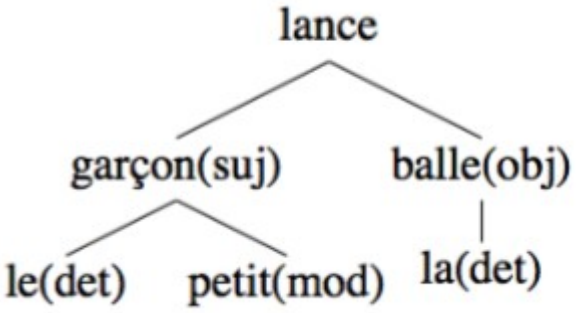
Analyse syntaxique

Objectif de l'analyse syntaxique

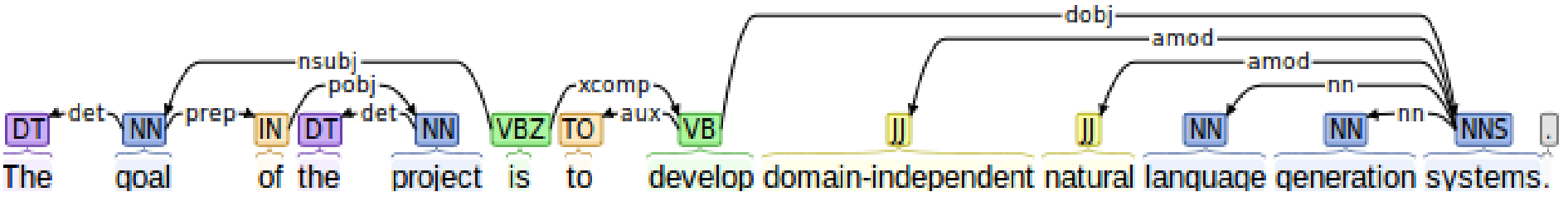
- Analyse syntaxique traditionnelle
  - Généralement fondée sur le paradigme génératif de Chomsky
  - Objet = générer tous et seulement les énoncés possibles dans une langue (énoncés grammaticaux)
  - En analyse = associer à un énoncé (phrase) grammatical(e) de la langue sa structure syntaxique
    - arbre des séquences de réécritures permettant d'obtenir la phrase à partir de l'axiome S de la grammaire

Exemple de sortie attendue



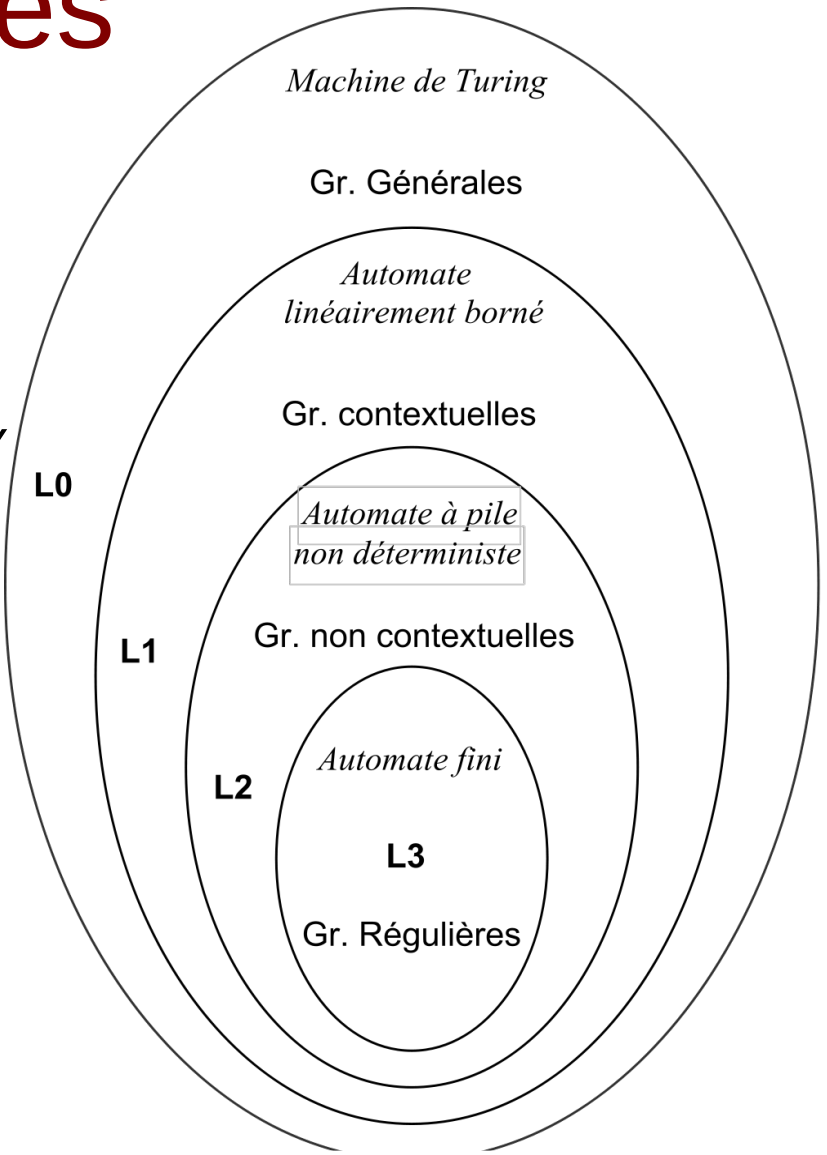


Exemple d'analyse en dépendances



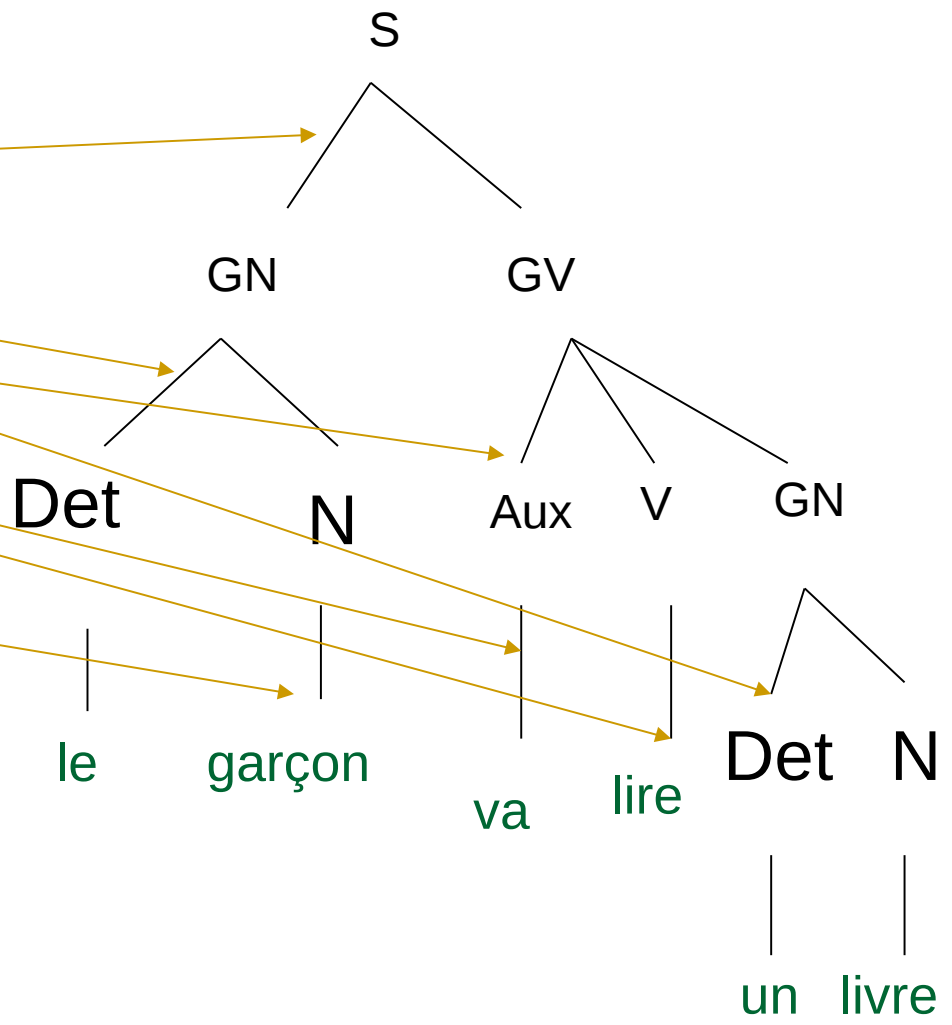
Grammaires

- G=(Vn, Vt, R, S)
  - Vn : vocabulaire non terminal
  - Vt : vocabulaire terminal
  - R : ensemble de règles de réécriture, X → Y
  - S : axiome de la grammaire
- Suivant les règles de R :
  - Grammaire non contrainte → trop « lâche »
  - Grammaire en contexte :
    - « X se réécrit Y dans le contexte u v »
    - uXv → uYv
  - Grammaire hors contexte : X → Y
  - Grammaire régulière (trop figée)
    - A → a ou A → aB



Grammaire hors-contexte

- Exemple :
- S → GN GV
- GN → Det N
- GV → (Aux) V GN
- Aux → va
- V→lire | bat |mange|…
- Det → le|la|les|un…
- N → garçon|livre|pomme…

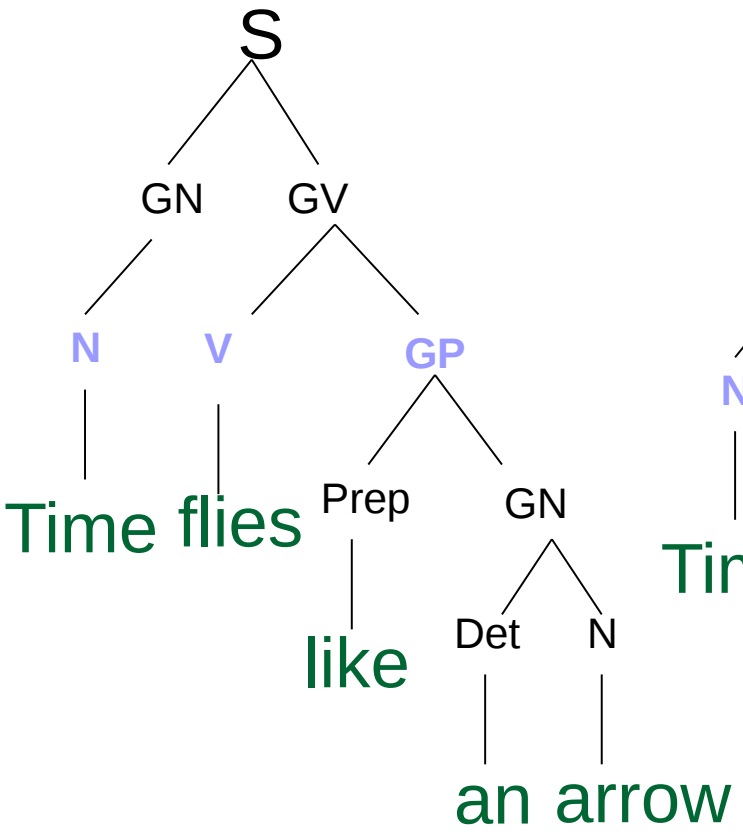


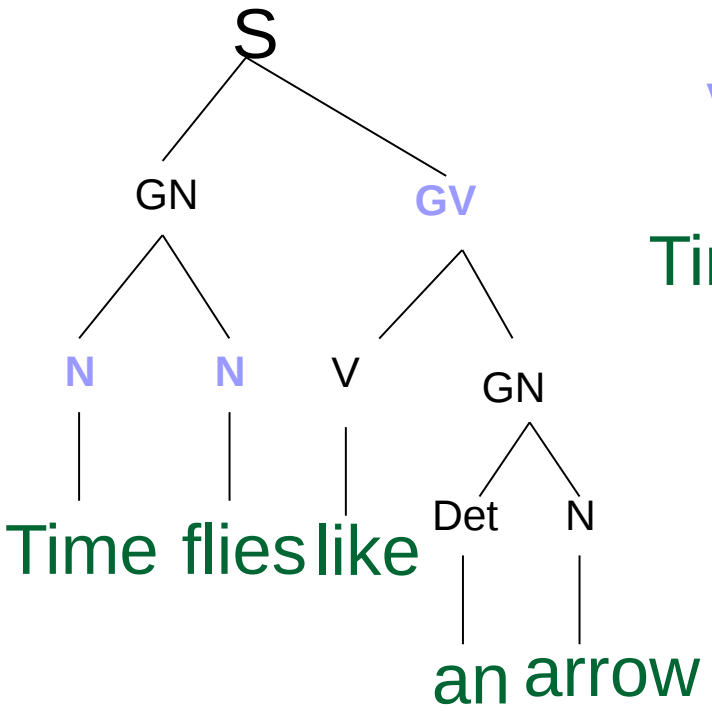
Le garçon va lire un livre

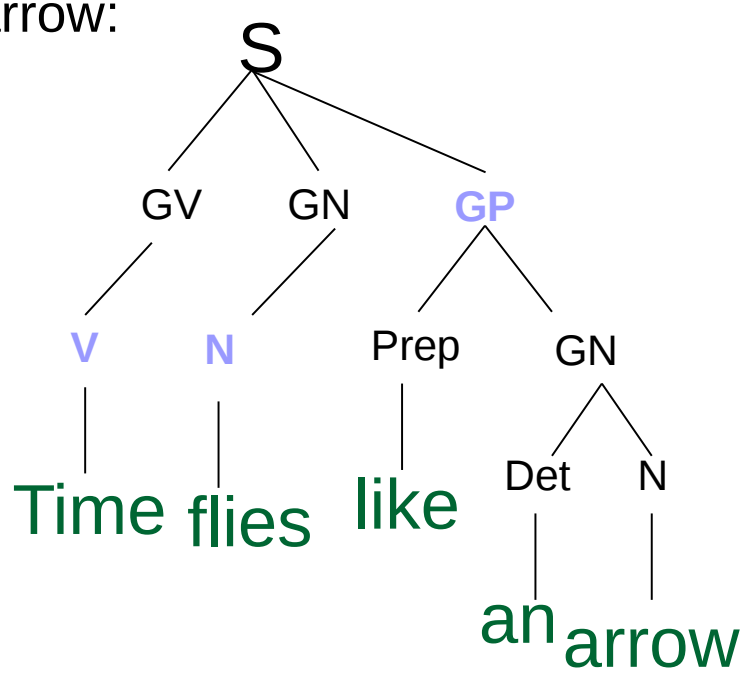
Mais aussi : *le pomme va mange la livre*

Grammaire hors contexte…

- Différences entre structure de surface et structures profondes
- Exemple « chomskyen » : Time flies like an arrow:







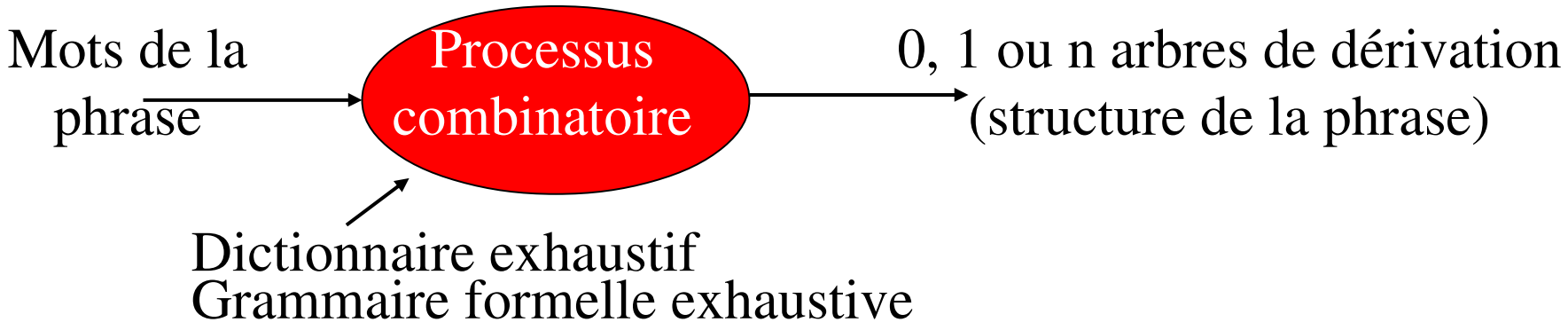
Analyse syntaxique traditionnelle

- Théorie des langages formels de Chomsky
  - Formalisation mathématique pas une théorie linguistique
  - La langue n'est pas un langage indépendant du contexte
    - Les accords
  - Grammaires contextuelles insuffisantes
    - Constituants discontinus : Combien cette salle a-t-elle de fenêtres ?

Exemples d'analyseurs

- Analyseurs fondés sur les formalismes des théories grammaticales
  - GPSG (Generalized Phrase Phrase-Structure Grammar, Gazdar et al 1985)
  - LFG (Lexical Functional Grammar, Kaplan & Bresnan 1982)
  - UCG (Unification Categorial Grammar, Clader et al 1988)
  - HPSG (Head-driven Phrase Structure Grammar, Pollard & Sag 1994)
- Autres
  - PATR : formalisme à structures de traits et unification
  - DCG (Definite Clause Grammar) : extension de Prolog

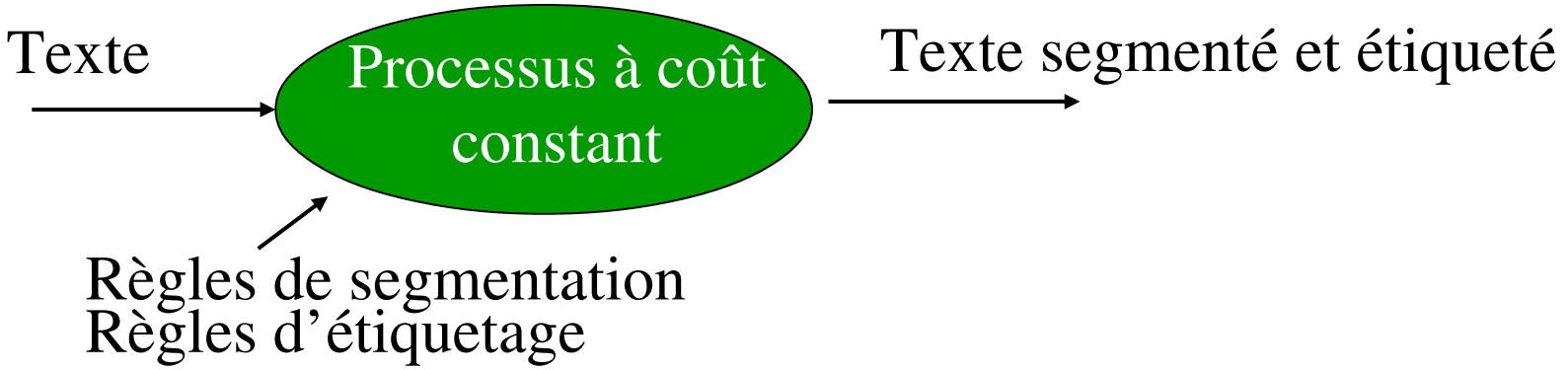
Analyse syntaxique traditionnelle



- Caractéristiques (HPSG, LFG, TAG, …) :
  - Règles de grammaire de type hors-contexte
  - Structures de traits
  - Unification
- Problème : manque de robustesse

Analyse robuste

Analyse robuste, analyse partielle, analyse de surface (shallow parsing)



- Approche empirique : héritage de la reconnaissance de la parole
- Travail sur texte réel, but opérationnel d'abord
- Analyse vue comme un processus informatique
- Principalement des méthodes statistiques

Notion de robustesse en TAL

- Robustesse : plusieurs définitions dans la littérature du TAL
- Idée commune :
  - Capacité d'un système de TAL à traiter des données linguistiques réelles (produites par des locuteurs indépendamment du système)
- Définition (pour un analyseur)
  - Capacité d'un système à produire des analyses utiles pour des textes réels
  - Analyses utiles : analyses (au moins partiellement) correctes et utilisables dans une tâche automatique (application)

(Salah Aït-Mokhtar, 2002)

Textes traités

- Texte réels : tels que produits par leurs auteurs
  - documentation technique
  - articles ou dépêches de presse
  - pages web, courrier électronique
  - sortie d'OCR
- Caractéristiques
  - aspects typographiques
  - mots inconnus, énoncés agrammaticaux
  - structures grammaticales particulières (ellipses, structures complexes, etc.)
  - expressions idiomatiques

Motivations derrière la robustesse

- Théorique
  - Confronter les modèles théoriques à des données réelles est une nécessité pour toute science empirique
- Pratique
  - Importants besoins d'applications capables de traiter des documents réels
  - Besoins favorisés par la disponibilité d'immenses quantités de documents électroniques (grâce notamment à Internet)

Propriétés nécessaires

- Une analyse au moins pour chaque entrée
  - Situations d'absence d'analyses fréquentes dans les analyseurs traditionnels
  - Enoncés agrammaticaux dans les textes réels
  - Mais, plus fréquemment : constructions grammaticales non prédites par le modèle ou les descriptions linguistiques de l'analyseur
- Nombre d'analyses concurrentes limité
  - Les analyseurs traditionnels produisent souvent de trop nombreuses analyses (parfois des milliers pour une longue phrase), dont des analyses redondantes (ambiguïtés artificielles)

Méthodes d’analyse robuste (1)

- Emergence de méthodes d’analyse robuste
  - Surtout à partir de la fin des années 80
- Trois tendances générales
  - Ajout de mécanismes ad hoc spécifiques pour rendre les analyseurs traditionnels robustes
  - Analyse à base de modèles statistiques
  - Analyse de surface à base de règles (rule-based shallow parsing)

Analyse de surface (*shallow parsing*)

- Idée de base
  - Limiter la « profondeur » et la richesse de l'analyse syntaxique
  - Prévoir la possibilité d'analyses partielles
- But
  - Obtenir des structures syntaxiques minimales, sous-spécifiées mais linguistiquement motivées (syntagme noyau = *chunk*)
  - Des structures utiles en tant que telles dans des applications
  - Première phase d'une analyse syntaxique plus complète

Exemple d’analyse

[Bill NP] [vit V] [l’homme NP] [sur la colline PP] [avec un télescope PP]

- Chunks : NP, V, PP
- Ambiguïté de rattachement implicite

Analyse de surface: étapes de traitements

- Prétraitement
  - Etiquetage morpho-syntaxique (segmentation, analyse morphologique, désambiguïsation)
- Analyse syntaxique de surface
  - Reconnaissance des syntagmes noyaux (chunks) : SN, SP, SV
  - Groupes complexes et propositions
  - Attribution de fonctions syntaxiques (Sujet, Objet, etc.)
- Analyse incrémentale

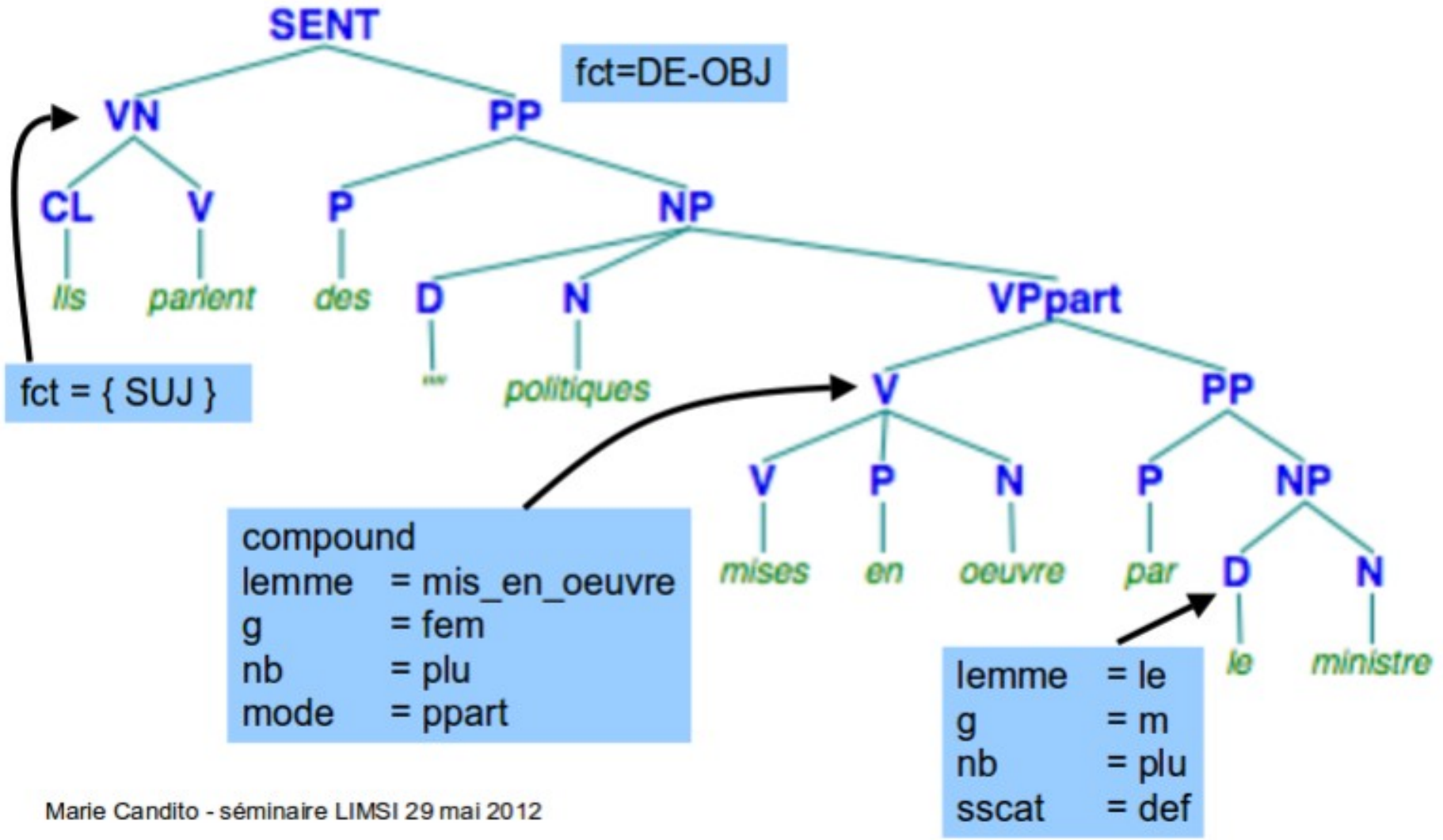
Analyse par apprentissage supervisé

- Nécessité de grands corpus annotés
  - Penn TreeBank pour l'anglais
  - French TreeBank pour le français

Corpus French TreeBank

- Projet initié en 1997
- corpus journalistique (Le Monde)
- 1 million de mots
- Annotations
  - Morphosyntaxiques
    - Pos
    - Sous-catégorisation
    - Inflection
    - Lemme
    - Parties pour mots composés
  - Constituants
  - Fonction

Représentation dans le FTB



Principe de l'analyse probabiliste en constituants

- Probabilistic context-free grammar (PCFG)
  - dès (Booth, 69)
  - une CFG + probabilités:
    - chaque règle est associée à une probabilité
    - probabilités telles que ∀ non terminal A : $\forall \sum_{\alpha : A \rightarrow \alpha \in G} P(A \rightarrow \alpha) = 1$
  - probabilité d'un arbre
    - = probabilité conjointe de toutes les applications de règles sous-jacente à l'arbre
    - «grammaires hors-contexte» $P(arbre) = \prod_{A \rightarrow \alpha \in arbre} P(A \rightarrow \alpha)$
    - => hypothèse d'indépendance entre chaque règle

(Candito, 2012)

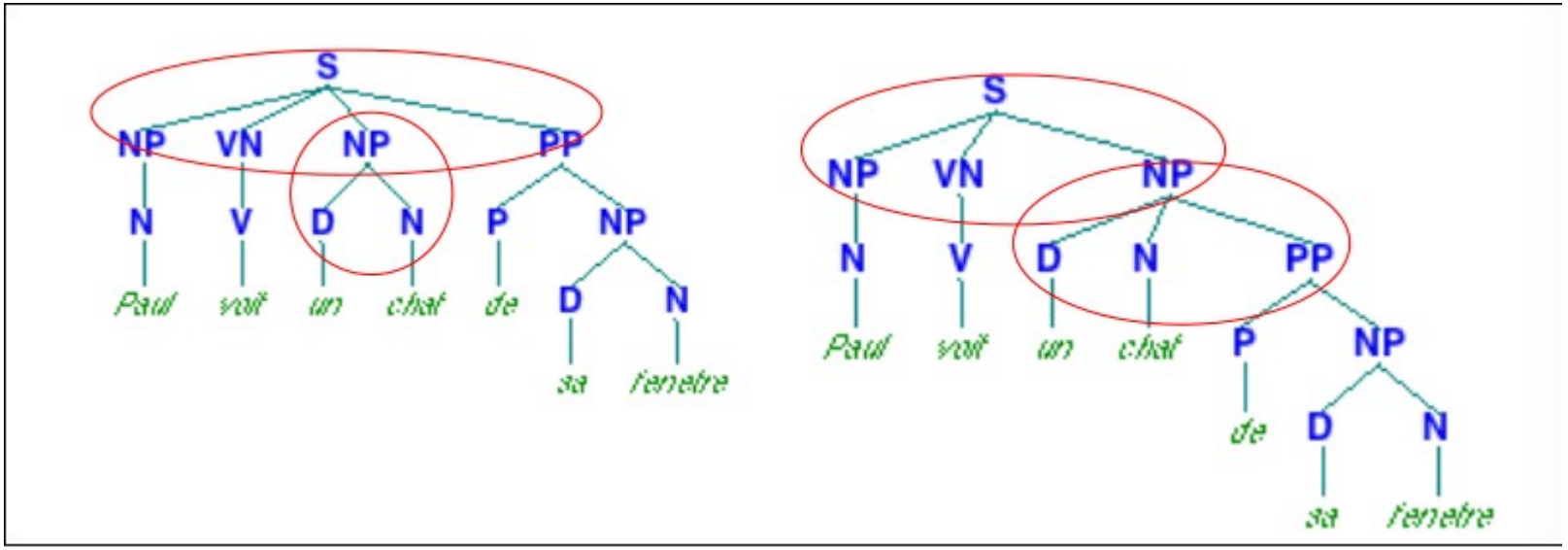
Extraire une PCFG d'un treebank

- CFG = règles rencontrés dans les arbres du corpus
- probabilités associées aux règles = estimées par fréquence relative (max de vraisemblance)

$$P(A \rightarrow \alpha) = \frac{C(A \rightarrow \alpha)}{\sum_{\beta : A \rightarrow \beta \in G} C(A \rightarrow \beta)}$$

Inconvénients des PCFG

- Hypothèses d'indépendance trop fortes
  - ne tiennent pas compte du lexique
    - par exemple, choix entre ces deux analyses

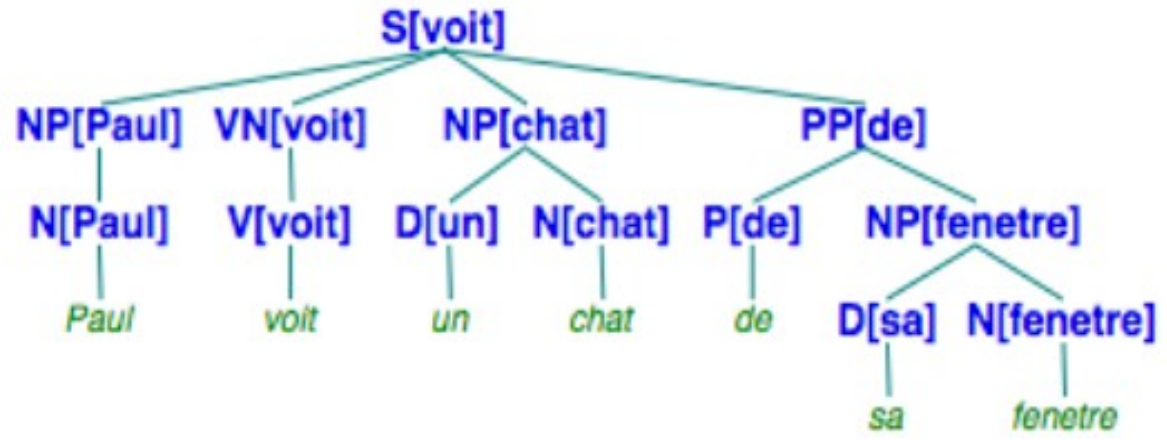


Inconvénients des PCFG

- il existe des dépendances structurelles entres les règles
  - exemple typique (Johnson, 98) sur Penn TreeBank
    - Les NP sujets ont plus de chance d'être pronominaux que nominaux, contrairement aux NP objets
    - P(NP -> PRO | (en sujet)) >> P(NP -> PRO| non sujet)
  - exemple en français
    - Syntagme adjectival : jamais prénominal si contient complément postadjectival
    - un [très charmant] garçon
    - *un [très charmant envers tous] garçon

Solution 1 : algorithme lexicalisé

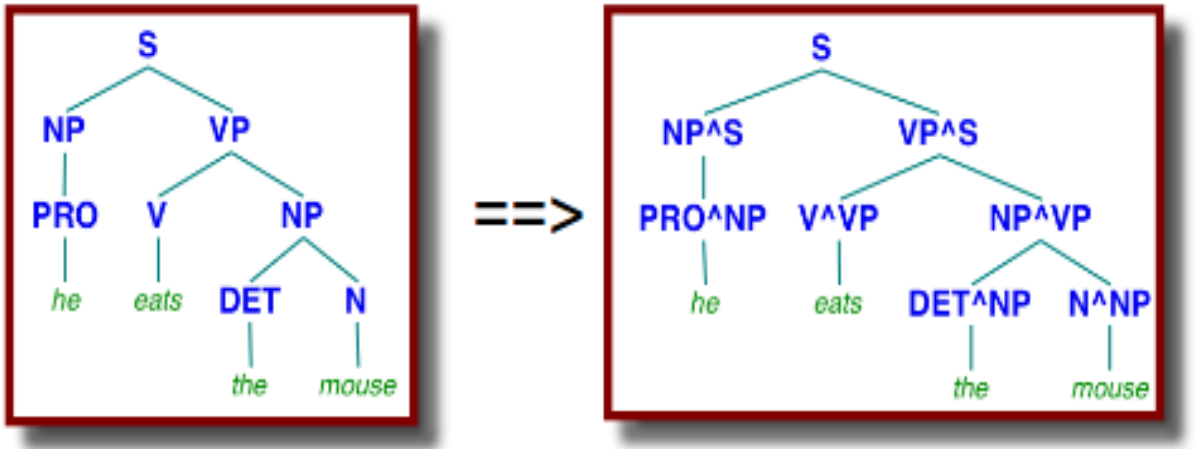
- Collins, 99 : lexicalisation des règles
  - tête lexicale associée à chaque règle d'un arbre



  - Probas sur règles augmentées :
  - S(voit,V) → NP(Paul,N) VN(voit,V) NP(chat,N) PP(de,P)

Solution 2 : division de symbole

- raffinage manuel des symboles non terminaux
  - Johnson, 98 : annotation par le nœud parent



- Klein et Manning, 03 : essais systématiques de raffinements intuitivement/linguistiquement intéressants
  - exemple : split du tag IN (pour prep) en 6 sous-catégories selon la prep
  - améliore les résultats
  - laborieux, splits dépendants du corpus

Solution 2bis : division automatique

- PCFG avec annotations latentes
  - trouver automatiquement les raffinages pertinents
    - (Matsuzaki et. Al. 2005), puis (Petrov, 06; 07)
- apprentissage
  - G0= la grammaire extraite directement du treebank
  - Créer itérativement G1...Gn comme suit :
    - DIVISION: Diviser tout symbole X de Gi-1 en 2 nouveaux symboles X1 et X2
      - règles de la forme Ax → By Cz
      - probabilités des règles avec annotations latentes sont estimées via une variante de l'algorithme EM
    - FUSION :Re-fusionner des paires de symboles dont la distinction s'avère inutile
      - ne garder que les divisions dont la fusion occasionnerait une forte perte de vraisemblance
    - LISSAGE : Lisser les probabilités de règles partageant le même symbole gauche
- très bons résultats
  - sur divers types de treebank

Analyse probabiliste en dépendances

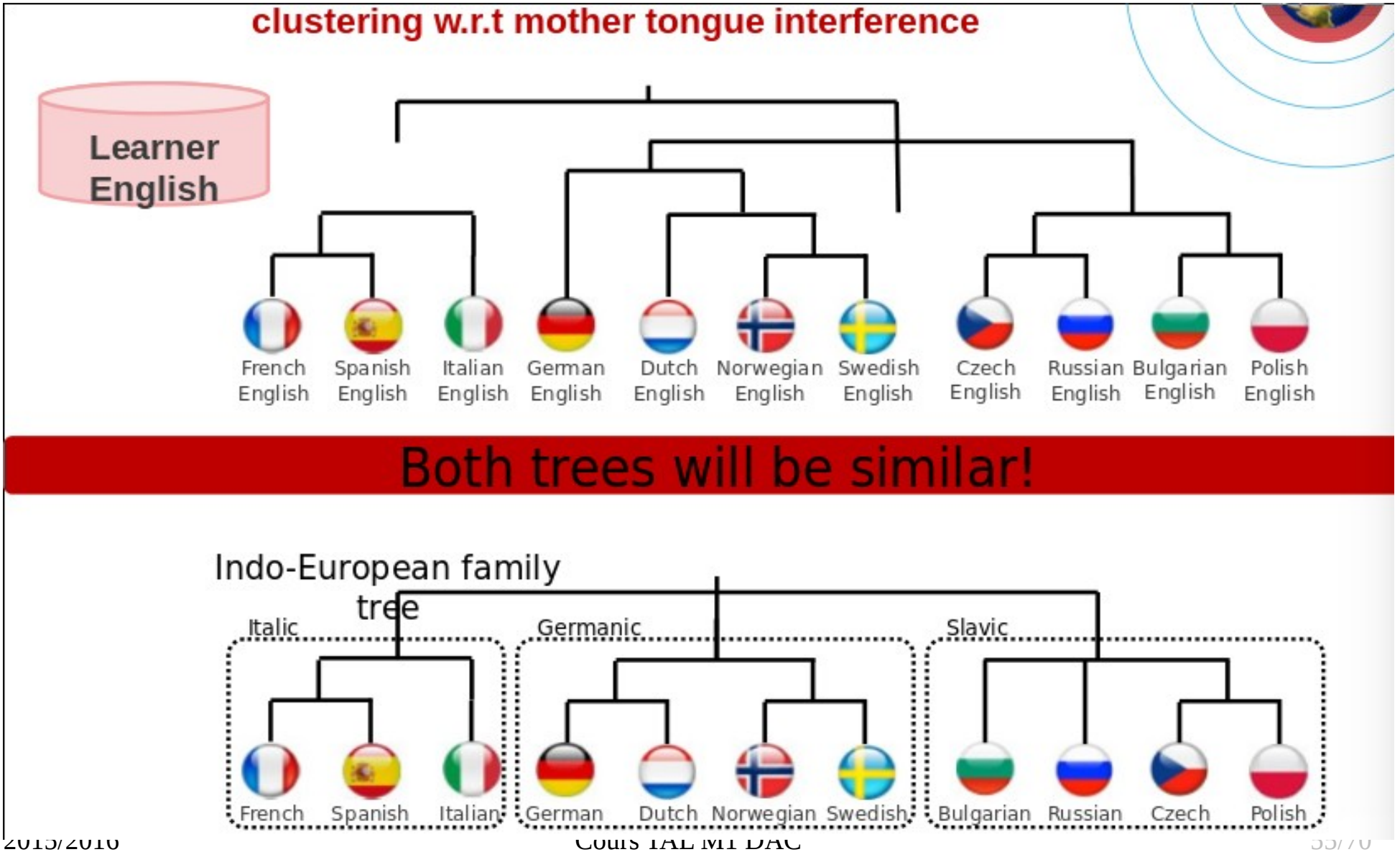
- Deux familles principales d'analyseurs
  - analyseurs à transitions
    - Yamada et Matsumoto, 2003; Nivre, 2003
    - implémentation de référence =MaltParser
  - basés sur les graphes
    - «Maximum spanning tree» parser McDonald, 2005
    - implémentation MSTParser, Bohnet...

Exemples d'applications

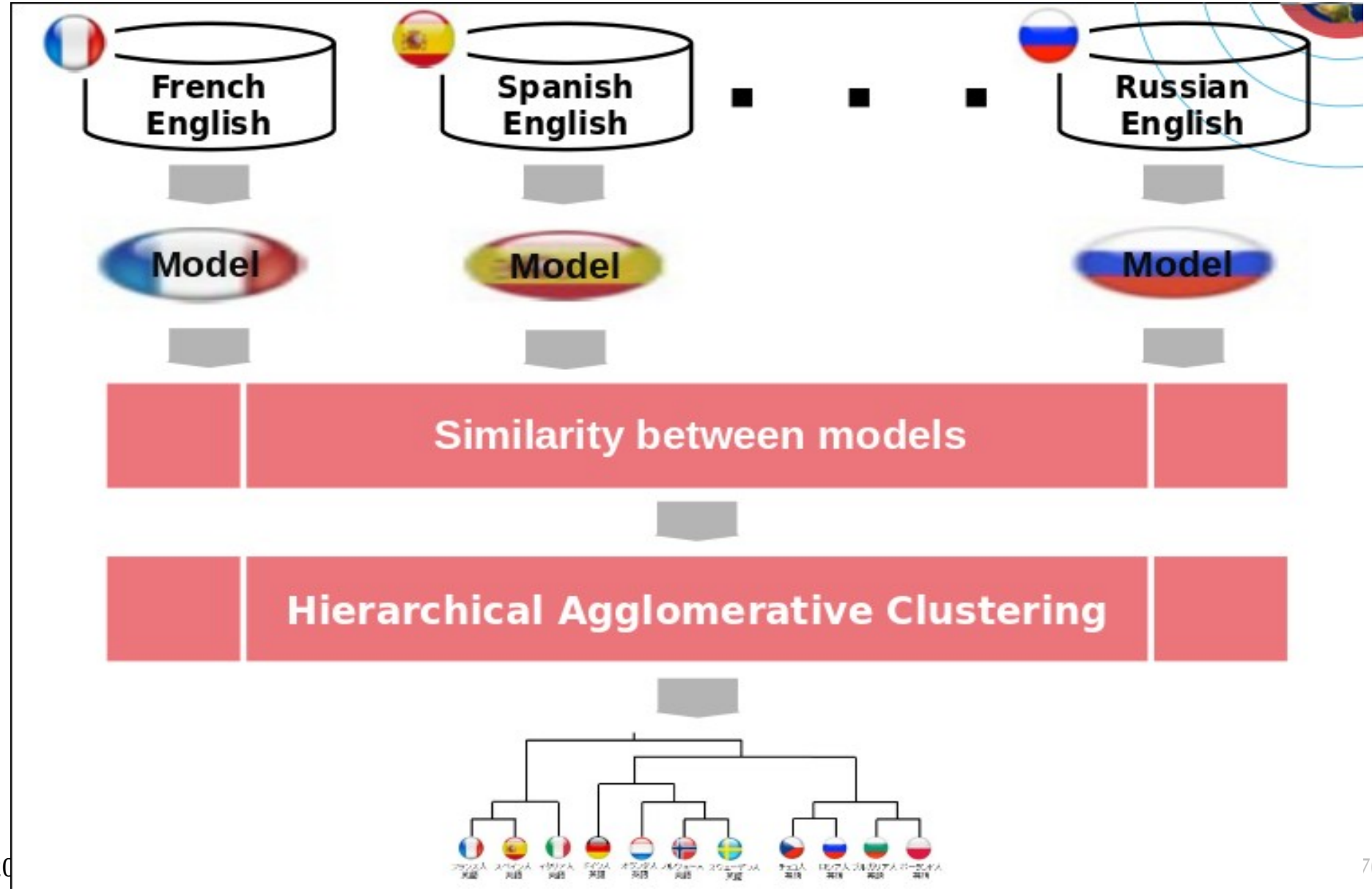
Application 1: retrouver les familles de langues par analyse de corpus

- Travaux de Ryo Nagata, 2012
- Idée : exploiter les interférences linguistiques de la langue maternelle en anglais
- par exemple
  - The alien wouldn't use my spaceship but **the hers**.
  - structure qui existe en français par exemple

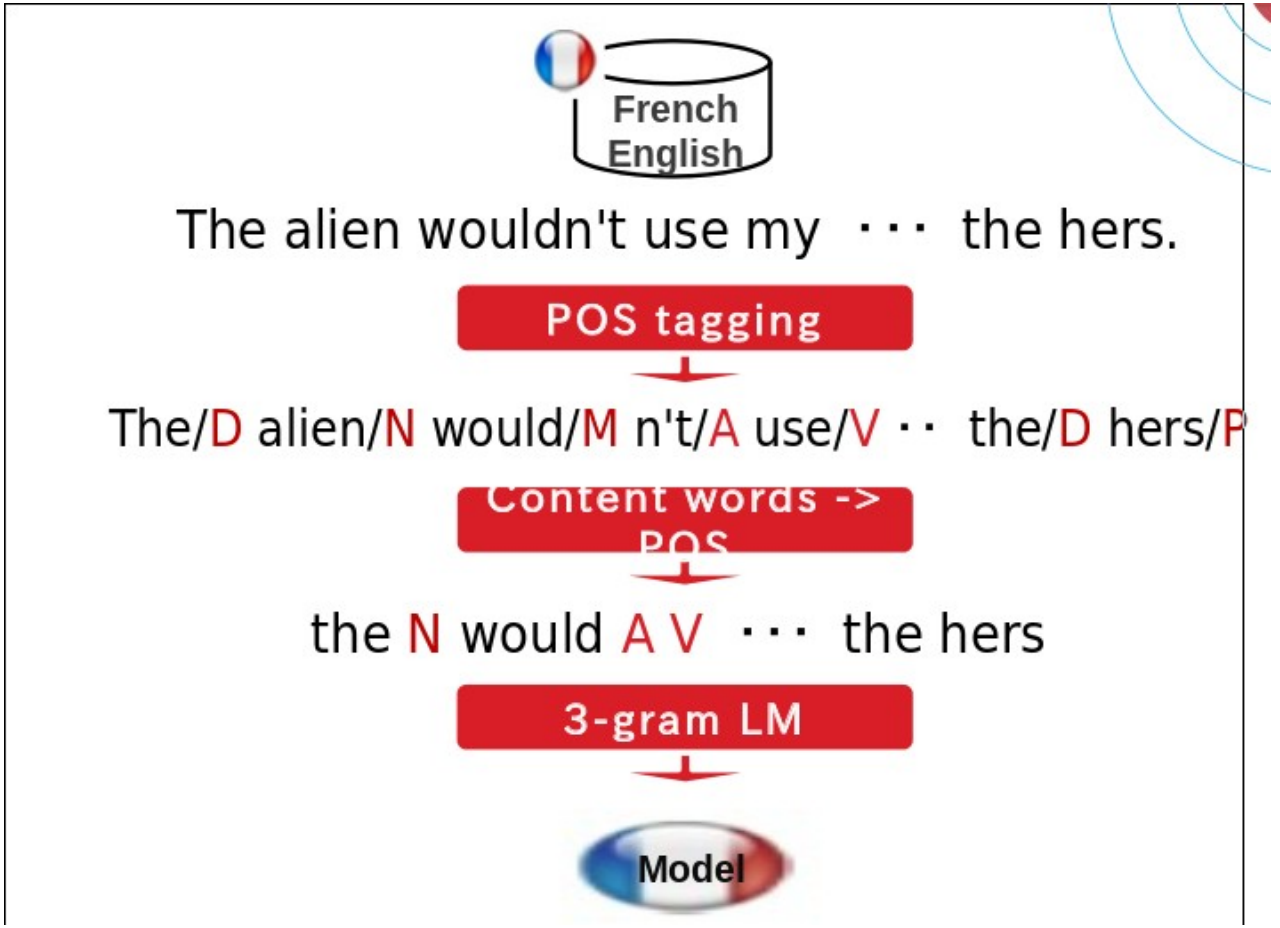
Hypothèse de recherche



Méthode


French English
Spanish English
Russian English
Model
Model
Model
Similarity between models
Hierarchical Agglomerative Clustering

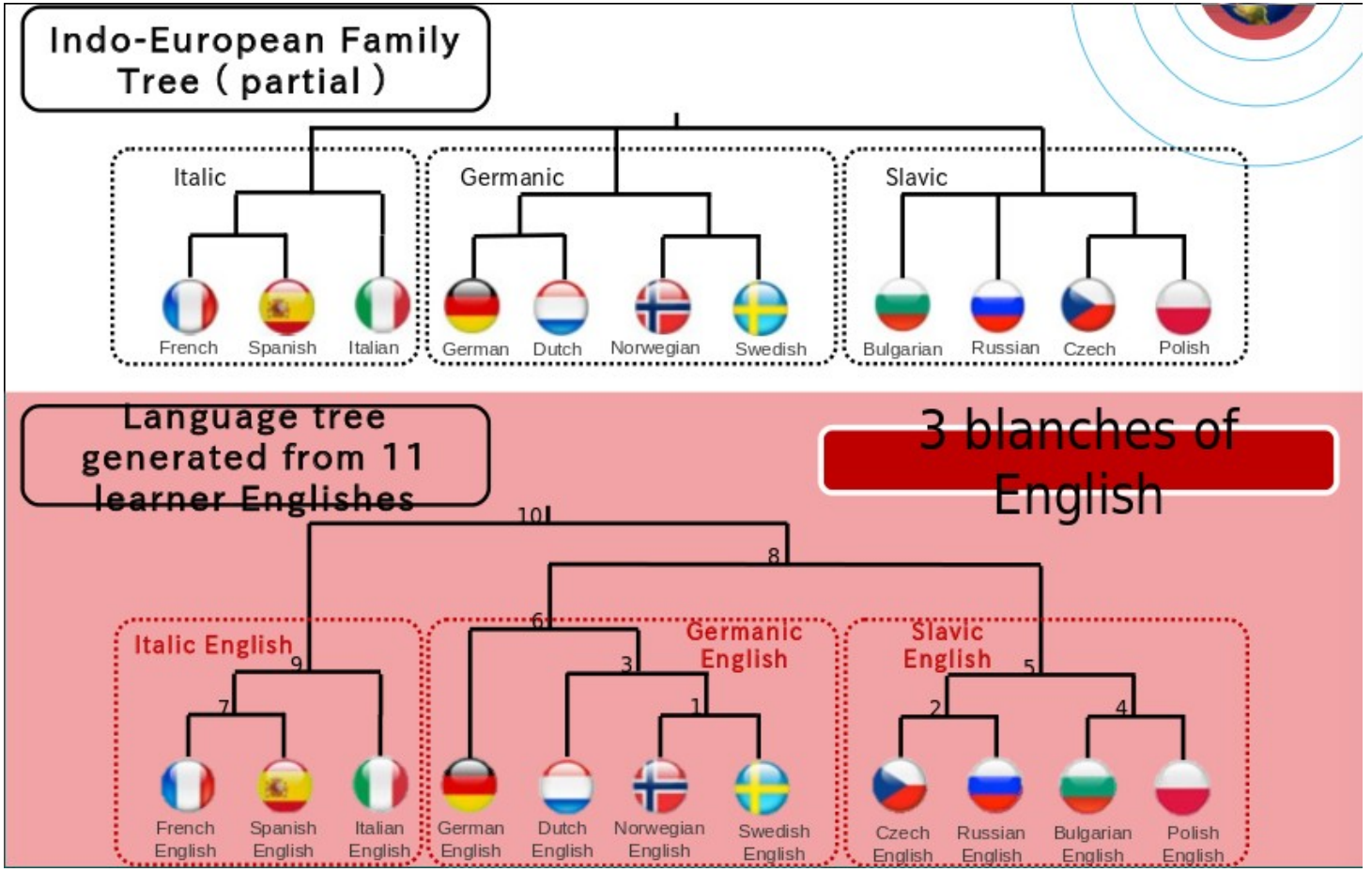
Création des modèles



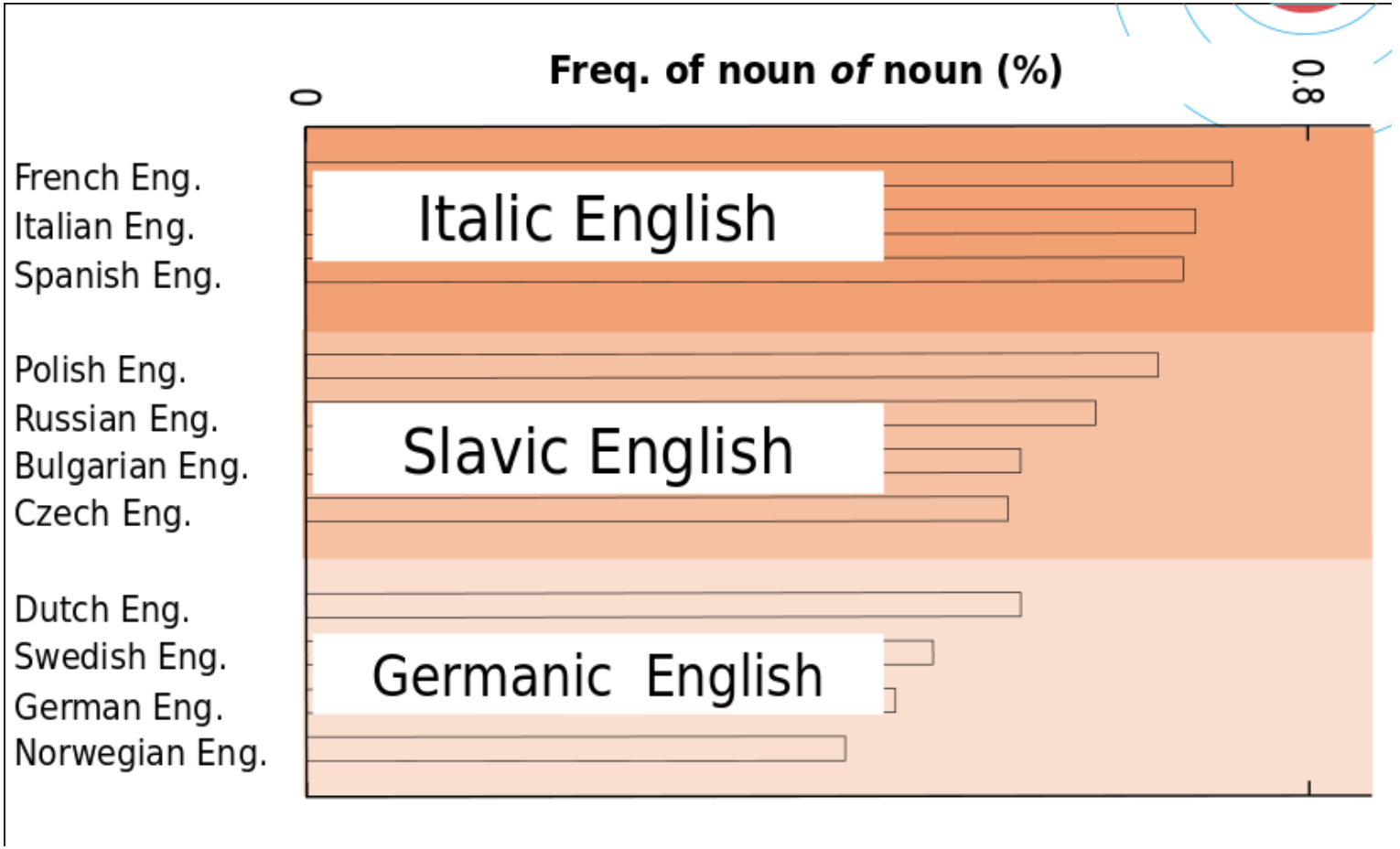
Expériences et résultats

- Corpus : ICLE corpus (Granger, 09)
  - corpus d'apprenants de l'anglais
  - 11 anglais
  - 20 millions de mots/ langue

Résultats


Indo-European Family Tree (partial)
Italic
Germanic
Slavic
French
Spanish
Italian
German
Dutch
Norwegian
Swedish
Bulgarian
Russian
Czech
Polish
Language tree generated from 11 learner Englishes
3 blanches of English
10
8
6
9
7
3
1
5
2
4
Italic English
Germanic English
Slavic English
French English
Spanish English
Italian English
German English
Dutch English
Norwegian English
Swedish English
Czech English
Russian English
Bulgarian English
Polish English

Analyse de résultats



English : cheese tart => noun of noun fréquent dans les langues romanes
Italic : tart of cheese (tarte au fromage)

Application 2 : IBM Watson

2011 : Watson joue à Jeopardy! contre les champions Brad Rutter and Ken Jennings… et gagne

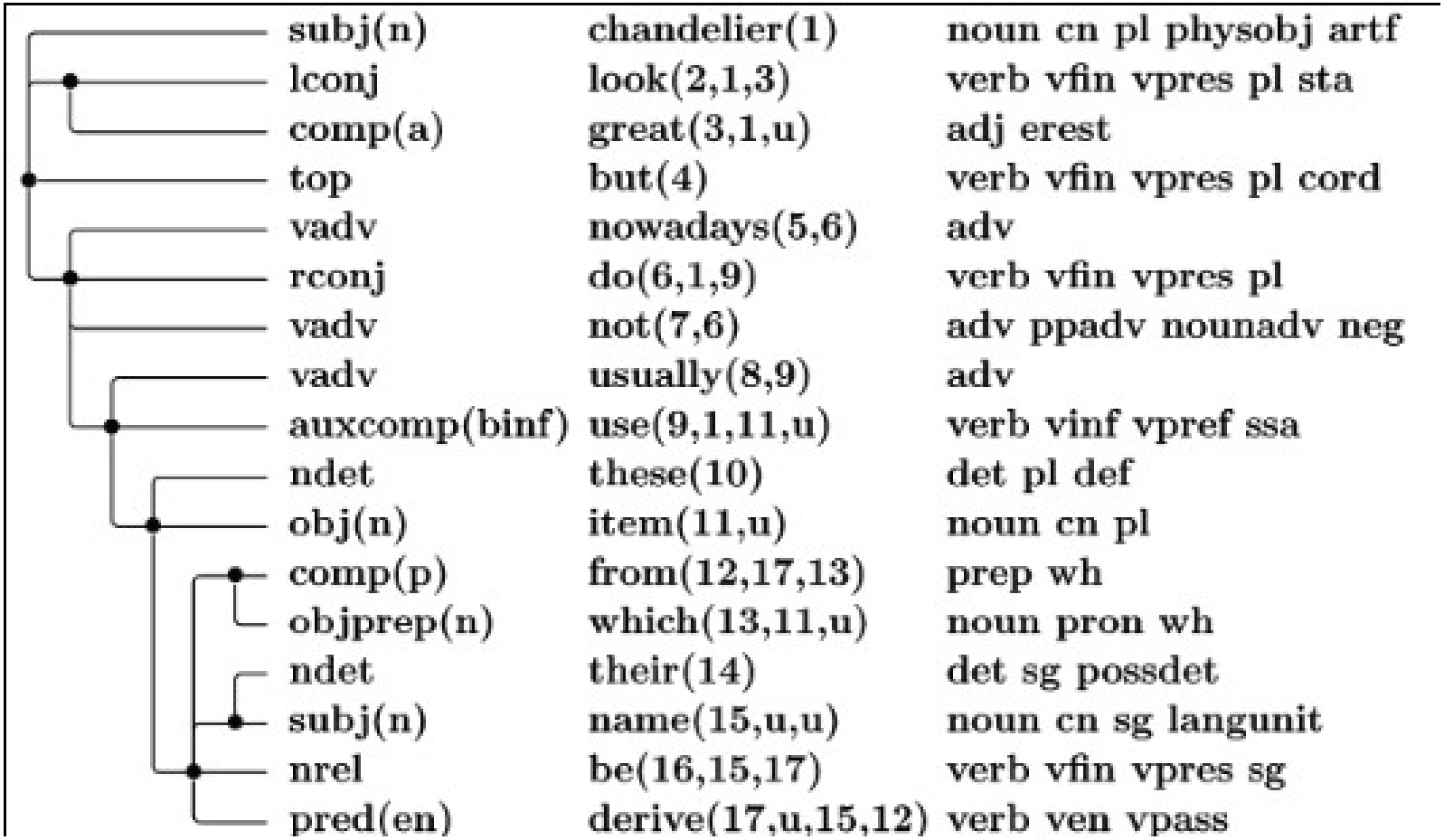
IBM Watson

- Programme d'intelligence artificielle capable de répondre à des questions en langage naturel
- Développé pour répondre à des questions du jeu Jeopardy !
  - In 1903, with presidential permission, Morris Michtom began marketing these toys
    - What are teddy bears?
  - This langage was invented in Warsaw in 1887 by Dr. L.L. Zamenhof
    - What is esperanto?
  - Cecilia Bartoli has unearthed & sung several forgotten arias by this « Four seasons » composer
    - Who is Antonio Vivaldi?

Analyse syntaxique dans Watson

- Analyse des questions et des documents grâce à une analyse syntaxique (et sémantique) profonde
  - English Slot Grammar (ESG) parser
    - produit un arbre syntaxique profond/ structure logique, ainsi qu'une structure de surface
  - Predicate Argument Structure (PAS) builder
    - simplifie l'analyse de l'ESG
      - supprime auxiliaires, voix passive = voix active...

Exemple d'analyse d'une question



subj(n)	chandelier(1)	noun cn pl physobj artf
lconj	look(2,1,3)	verb vfin vpres pl sta
comp(a)	great(3,1,u)	adj erest
top	but(4)	verb vfin vpres pl cord
vadv	nowadays(5,6)	adv
rconj	do(6,1,9)	verb vfin vpres pl
vadv	not(7,6)	adv ppadv nounadv neg
vadv	usually(8,9)	adv
auxcomp(binf)	use(9,1,11,u)	verb vinf vpref ssa
ndet	these(10)	det pl def
obj(n)	item(11,u)	noun cn pl
comp(p)	from(12,17,13)	prep wh
objprep(n)	which(13,11,u)	noun pron wh
ndet	their(14)	det sg possdet
subj(n)	name(15,u,u)	noun cn sg langunit
nrel	be(16,15,17)	verb vfin vpres sg
pred(en)	derive(17,u,15,12)	verb ven vpass

Application 3 : simplification automatique de textes

- Travaux de (Brouwers et al., 2011 ; 2014)
- Contexte : capacité de lire rapidement et efficacement = atout important mais pas toujours maîtrisé
  - problème : trop grande complexité des textes
- Problématique :
  - simplification automatique = rendre des textes plus abordables tout en garantissant l'intégrité de leur contenu et en veillant à en respecter la structure
    - Chuck Bartowski est un nerd, un passionné d'ordinateurs qui travaille au Buy More de Burbank.
    - Chuck Bartowski est un passionné d'ordinateurs. Il travaille au Buy More de Burbank

Constitution d'un corpus

- À partir de Wikipédia et Vikidia
  - alignement de phrases
    - Wikipédia : Un archipel est un ensemble d’îles relativement proches les unes des autres. Le terme «archipel» vient du grec ancien "Archipelagos", littéralement «mer principale» (de "archi" : «principal» et "pélagos" : «la haute mer»). En effet, ce mot désignait originellement la mer Égée, caractérisée par son grand nombre d’îles (les Cyclades, les Sporades, Salamine, Eubée, Samothrace, Lemnos, Samos, Lesbos, Chios, Rhodes, etc.).
    - Vikidia : Un archipel est un ensemble de plusieurs îles, proches les unes des autres. Le mot «archipel» vient du grec "archipelagos", qui signifie littéralement «mer principale» et désignait à l’origine la mer Égée, caractérisée par son grand nombre d’îles.

Typologie de simplifications

- Lexicales
  - synonymes ou hyperonymes : située dans le land → en
  - références anaphoriques plus explicites : il → l'homme
  - utilisation d'une définition ou paraphrase au lieu d'un mot complexe
  - traductions : Estado novo → État nouveau
- Discours
  - inversions de propositions : le général avant le précis
    - Antoine Marie Jean-Baptiste Roger de Saint- Exupéry, né le 29 juin 1900 à Lyon et disparu en vol le 31 juillet 1944, Mort pour la France, est un écrivain, poète et aviateur français.

      → Antoine de Saint-Exupéry, né le 29 juin 1900 à Lyon, mort le 31 juillet 1944 disparu en vol, était un écrivain et aviateur français.
  - ajout d'informations, d'explications, d'exemples
  - suppression d'informations secondaires à la compréhension

Typologie de simplifications

- Syntaxiques
  - temps familiers plutôt que littéraires : rencontra → rencontre
  - suppressions : compléments circonstanciels, adverbes…
    - Il est peuplé de 710 231 habitants, soit l'équivalent du département français du Gard.

      → Il y a environ 686 293 habitants.
  - modifications (mise entre parenthèses, déplacement de proposition ou complément circonstanciel, structure clivée → non clivée, proposition secondaire → principale, etc.)
    - C'est à Aix qu'arriva en 802 l'éléphant blanc.

      → En 802 arriva à Aix un éléphant blanc.
  - divisions
  - regroupements

Méthode

- 19 règles de simplification syntaxique
- Application des règles
  - repérage des structures par expressions régulières sur les arbres
  - transformations
  - application récursive
  - sélection par Integer Linear Programming selon plusieurs critères de lisibilité :
    - longueur de la phrase
    - longueur des mots
    - familiarité du vocabulaire
    - présence de termes clés
- Résultats : environ 80 % de phrases grammaticalement correctes

Quelques liens

- Étiqueteurs morpho-syntaxiques
  - TreeTagger (en et fr)
    - http://nlp.stanford.edu/software/corenlp.shtml
  - Melt (fr)
    - https://gforge.inria.fr/projects/lingwb
- Analyseurs syntaxiques
  - BONSAI (fr)
    - http://alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html
  - Stanford parser (en et fr)
    - http://nlp.stanford.edu/software/lex-parser.shtml
- Toolkit
  - OpenNLP (modèles en)
    - https://opennlp.apache.org/
  - Stanford CoreNLP (modèles en)
    - http://nlp.stanford.edu/software/corenlp.shtml