L'ordinateur face à la richesse des langues

Éric de la Clergerie
<Eric.De_La_Clergerie@inria.fr>

http://alpage.inria.fr
INRIA Paris-Rocquencourt / Univ. Paris Diderot

Rencontre ISN
Rocquencourt – 4 Décembre 2013

Le propre de l'homme ?

Très grande diversité au travers d'au moins 6000 langues dans le monde, dont les langues des signes

Langage et ordinateur

Un vieux rêve: créer une intelligence artificielle

Test de Turing (1950): repose sur une conversation entre un humain et un programme
$\Longrightarrow$ maîtrise du langage

ELIZA (Weizenbaum 1966, ancêtre des agents conversationnels [*chatbots*])

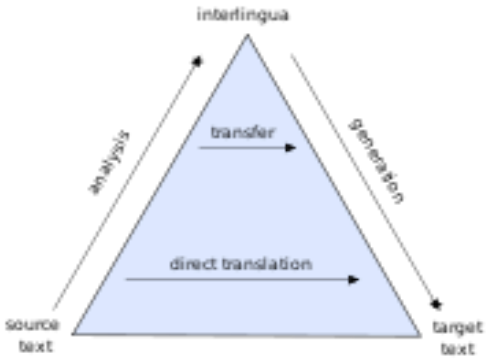
```
I am the psychotherapist.  Please, describe your problems.
   I'm not feeling well
Why do you say ``i'm not feeling well''?
   Well, I've no energy left
Is it because of your plans that you say ``well you have no
   All my plans are total failures
Maybe your life has something to do with this.
```

Chaque année depuis 1991, le prix Loebner récompense les meilleurs *chatbots*

Traduction automatique: déjà une longue histoire

Les premiers travaux sur le traitement du langage initiés pour des tâches de traduction (contexte de la guerre froide)

- première démonstration dès 1954 par IBM (russe -> anglais)
- coup de frein suite au rapport de Y. Bar-Hillel en 1960, préconisant des outils d'aide à la traduction
- développement de plusieurs générations d'approches et outils



- approche directe (mots à mots)
- par transfert (arbre syntaxique source vers arbre syntaxique cible)
- par interlangue pivot (représentation sémantique non liée à une langue)

De nos jours, traduction statistique: GOOGLE TRANSLATE

c'	est	à son ami qu'il	devrait parler
this	is	his friend that he	should talk

Siri, dois-je prendre mon parapluie ?

http://www.youtube.com/watch?v=xIBezLFLjiI

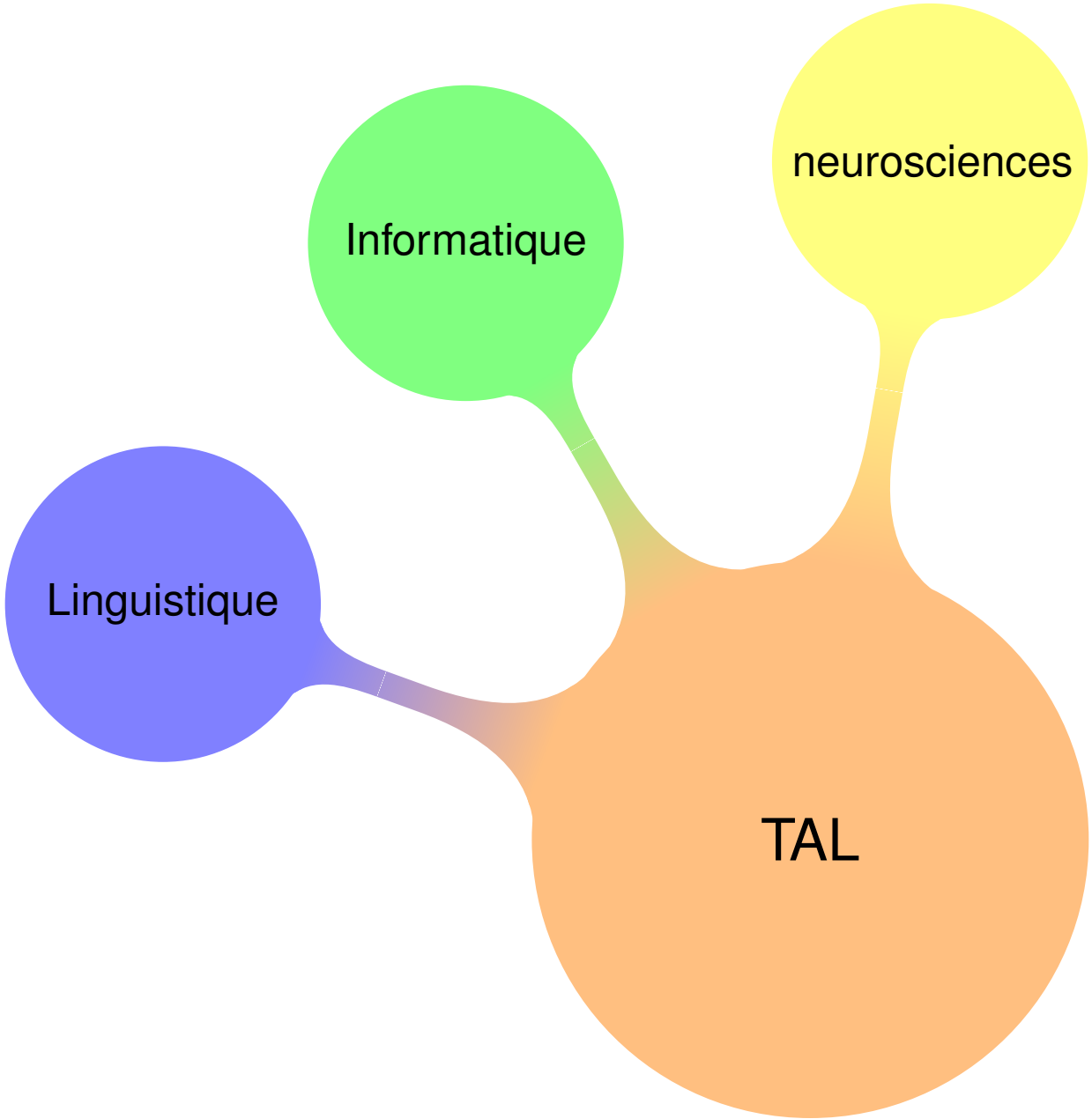
L'assistant vocal SIRI d'Apple se met en quatre pour vous servir !
(mais voir aussi http://www.youtube.com/watch?v=WGxDaX1__yI)

Et la réponse est ? . . . Élémentaire, mon cher Watson !

http://www.youtube.com/watch?v=WFR3lOm_xhE

WATSON est un programme développé par IBM (et un superordinateur !) champion du jeu télévisé *Jeopardy*

TAL, quesako ?



Niveaux de traitement

Paul, je t'ai dit que François Flore est sorti furieux de chez son banquier car celui-ci lui avait ex abrupto refusé son prêt pour sa future maison ?

Pragmatique: contexte & connaissances
référants: celui-ci=banquier, lui=son=sa=François, t'=Paul
structures argumentatives: refus explique furieux
scénarios, implicites

Sémantique: sens des énoncés et des mots
structure prédicatives, rôle des actants (agent, patient, ...)
`refuser(agent=celui-ci,patient=lui,theme=prêt)`

Syntaxe: structure des phrases et relations entre mots
fonctions syntaxiques (sujet, objet, ...) : **celui-ci**=sujet, **prêt**=objet, **lui**=obj indirect de **refusé**

Morphologie: les mots et leur structure (**lubéronisation**)
découpage du texte en mots, catégories syntaxiques:
`celui/pro -ci/adj lui/cld avait/aux ex_abrupto/adv ...`
flexion (conjugaison) : **avait**=`avoir+3s+Ind+Imparfait`
entitées nommées (personnes, lieux, ...) : (François Flore) `PERSON_m`

Quelles applications ?

De nombreuses applications potentielles, dans les laboratoires de recherche et dans quelques produits:

- correction orthographique, grammaticale, stylistique (CORDIAL, WORD, ...)
- recherche d'information
- fouille de textes, acquisition de connaissances
- fouille d'opinions (e-reputation)
- extraction d'informations et systèmes de Questions-Réponses (WATSON),
- traduction (GOOGLE TRANSLATE, SYSTRAN, MOSES, ...) et aide à la traduction
- résumé automatique
- génération
- dialogue Homme-Machine (SIRI), agent conversationnels (ELIZA, ALICE)
- reconnaissance vocale, dictée vocale (NUANCE)
- synthèse vocale
- ...

Quelles difficultés ?

La variabilité

- plusieurs manières d'exprimer une même idée

  There's more than one way to do it (Perl slogan, Larry Wall)

- choix des mots & diversité des constructions linguistiques
- des construction contrôlées par les mots, leur sens, le contexte, l'intonation, . . .

L'ambiguïté du langage, à tous les niveaux

- morphologique: **des** = **de les** ou (det. indéfini pluriel)
  la belle ferme le voile
- syntaxique
- sémantique: **avocat** = *homme de loi* ou *fruit* (*avocat véreux*)
  Paul propose à Jean de passer : qui passe ?
- . . .

Focus sur le niveau syntaxique pour ces deux problématiques

Analyse syntaxique

Objectif: retourner la structure grammaticale d'une phrase, en particulier les relations existantes entre les mots.

Pour cela, on peut s'appuyer sur:

- une grammaire formelle ($\not\sim$ grammaire scolaire)
  une collection de structures partielles d'analyse pour les divers phénomènes syntaxiques
- des règles du jeu pour combiner les structures de la grammaire

En pratique, une grande
variété de types de grammaires, reflétant des formalismes
et théories linguistiques divers (TAG, LFG, HPSG, CCG, . . .).

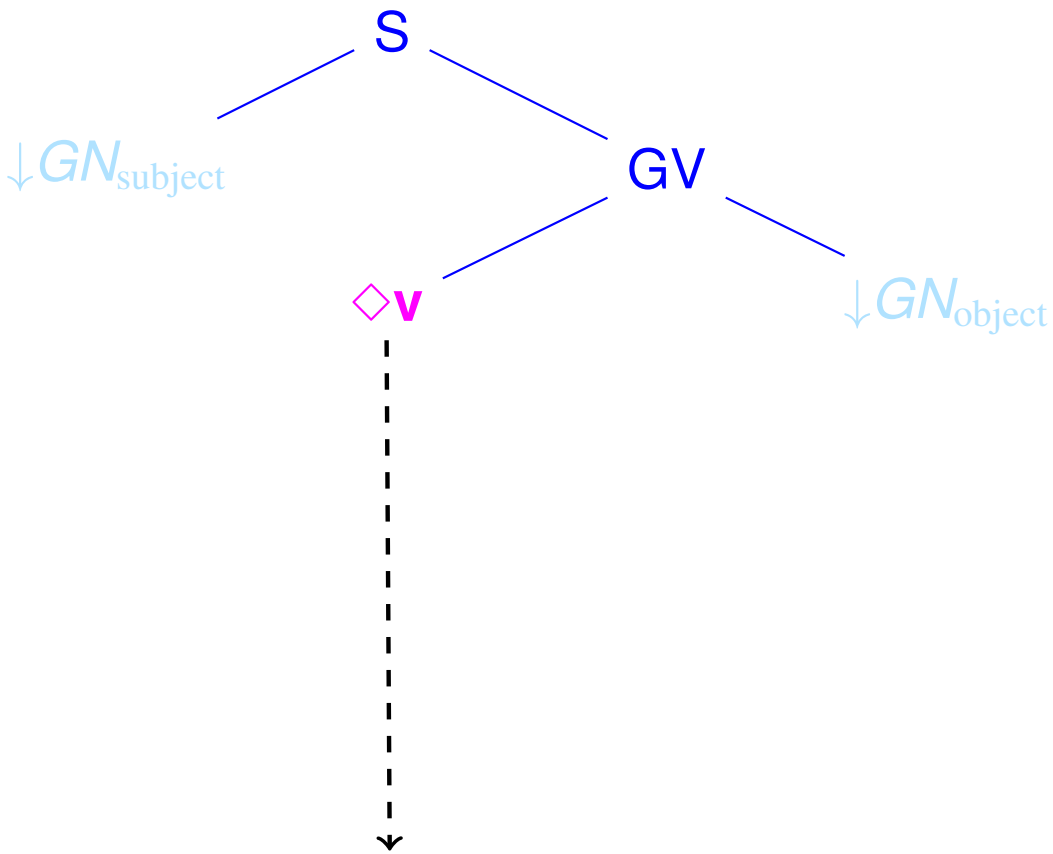
Développement par ALPAGE de FRMG pour le français,
en tant que *Grammaire d'Arbres Adjoints* (TAG Joshi)
en ligne sur `http://alpage.inria.fr/parserdemo`


Aravind Joshi

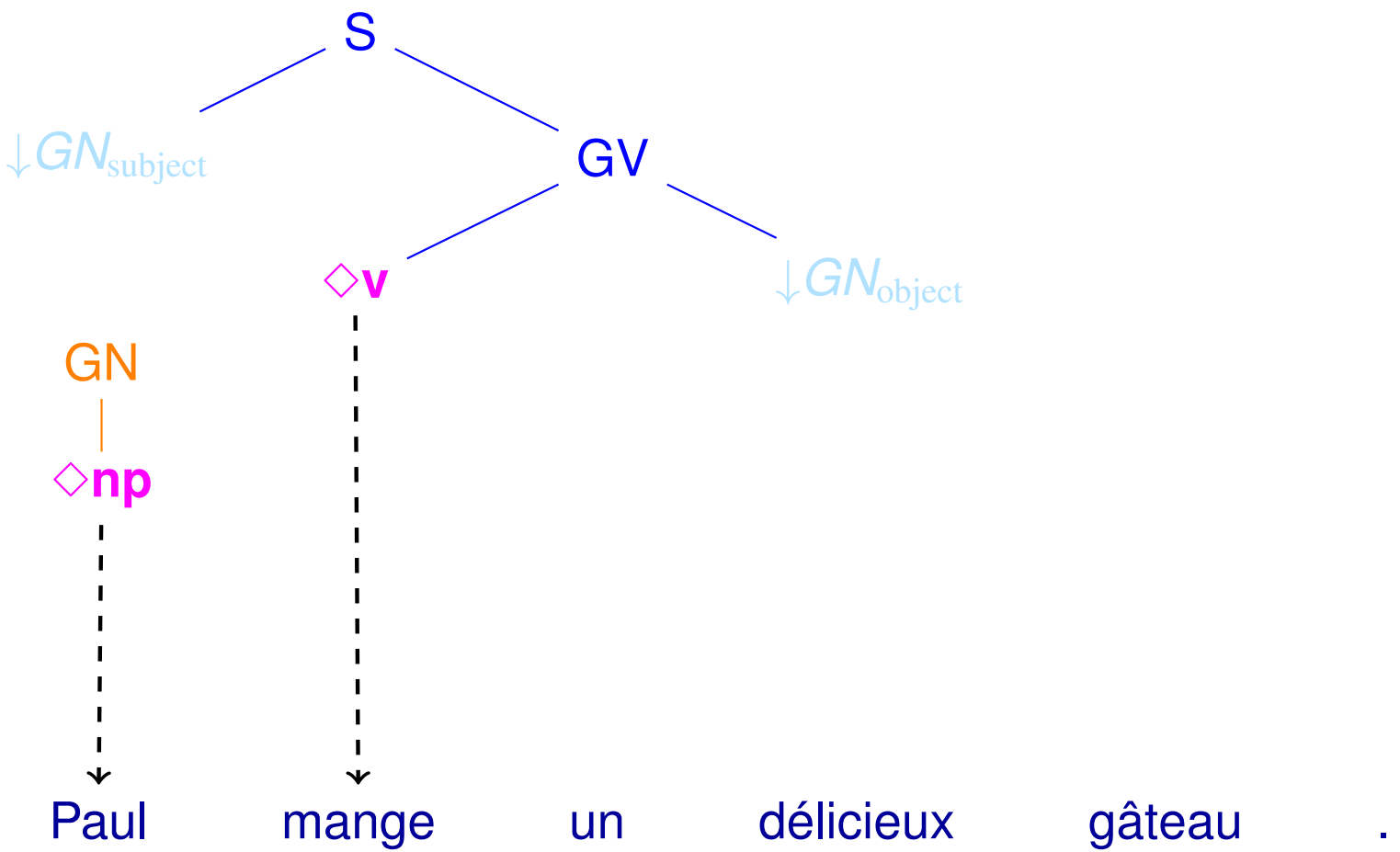
Exemple d'analyse

Paul mange un délicieux gâteau .

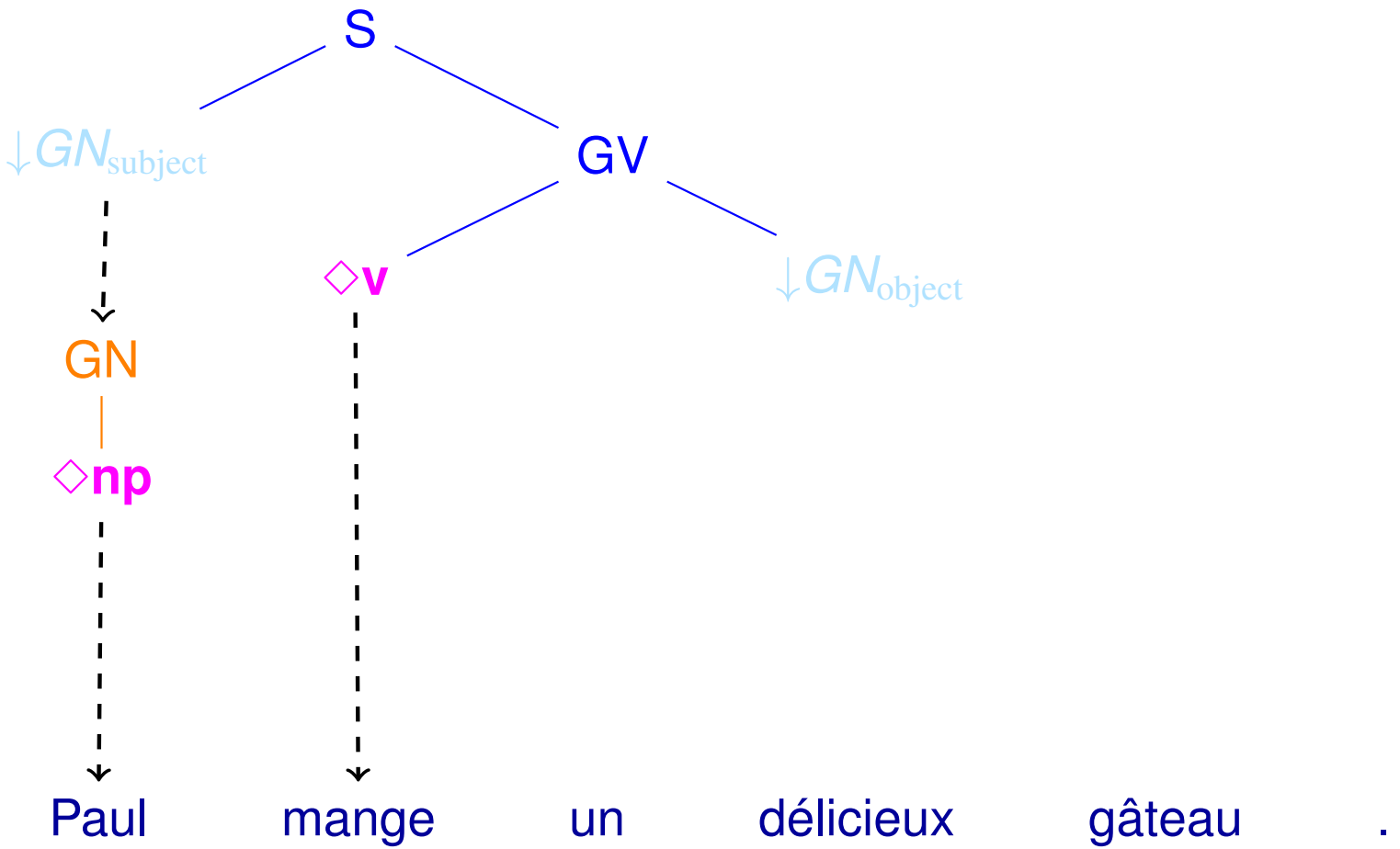
Exemple d'analyse



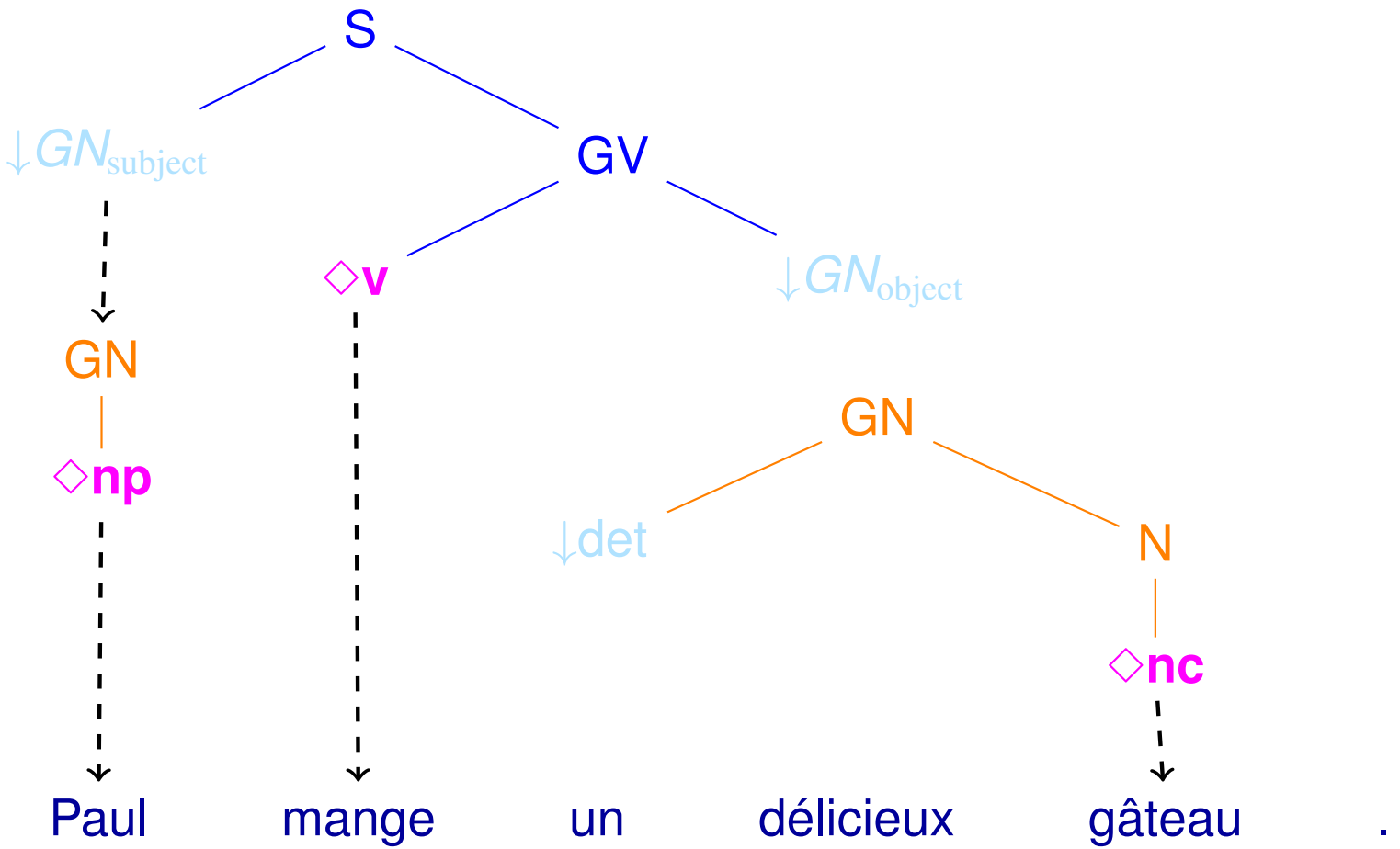
Exemple d'analyse



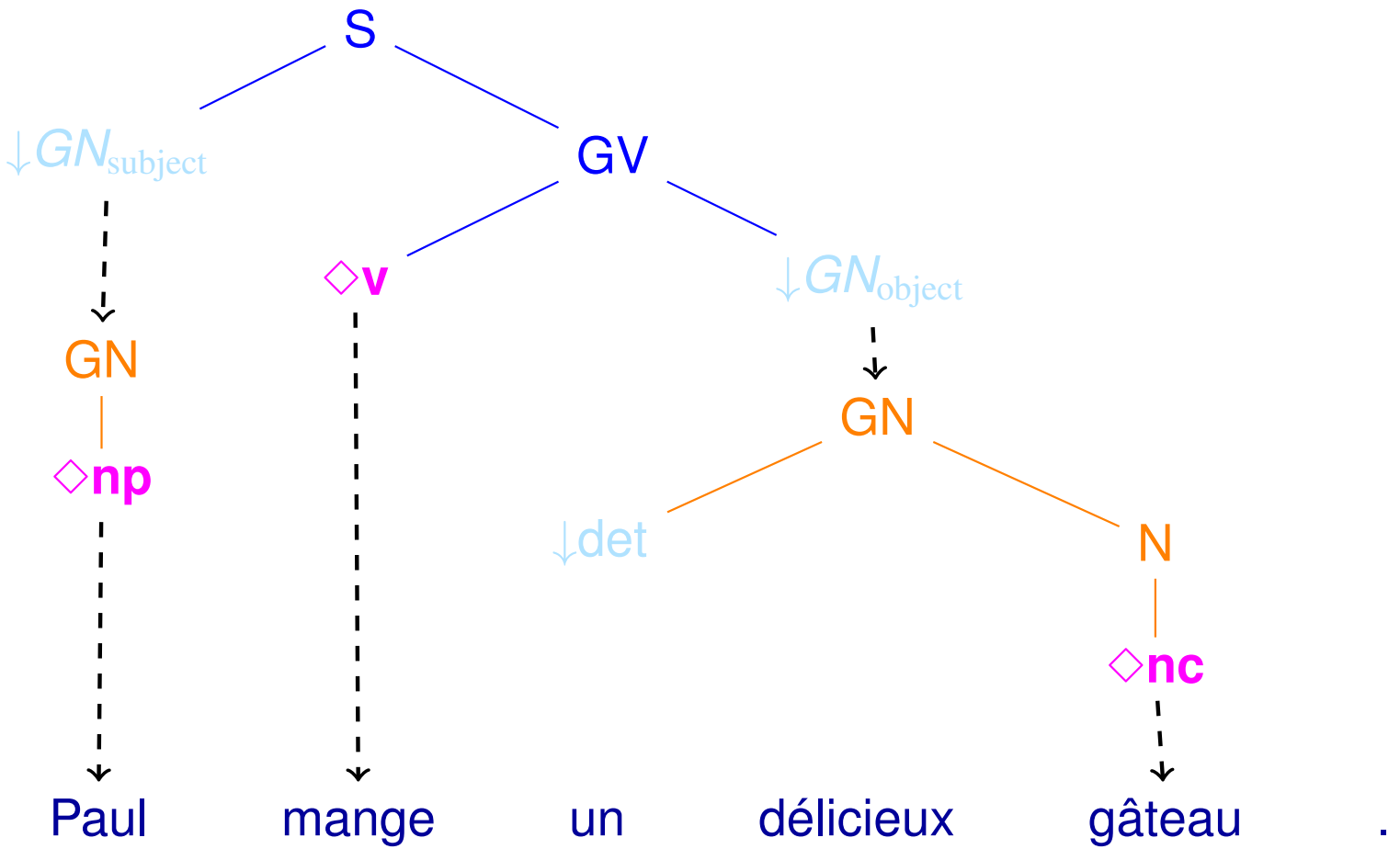
Exemple d'analyse



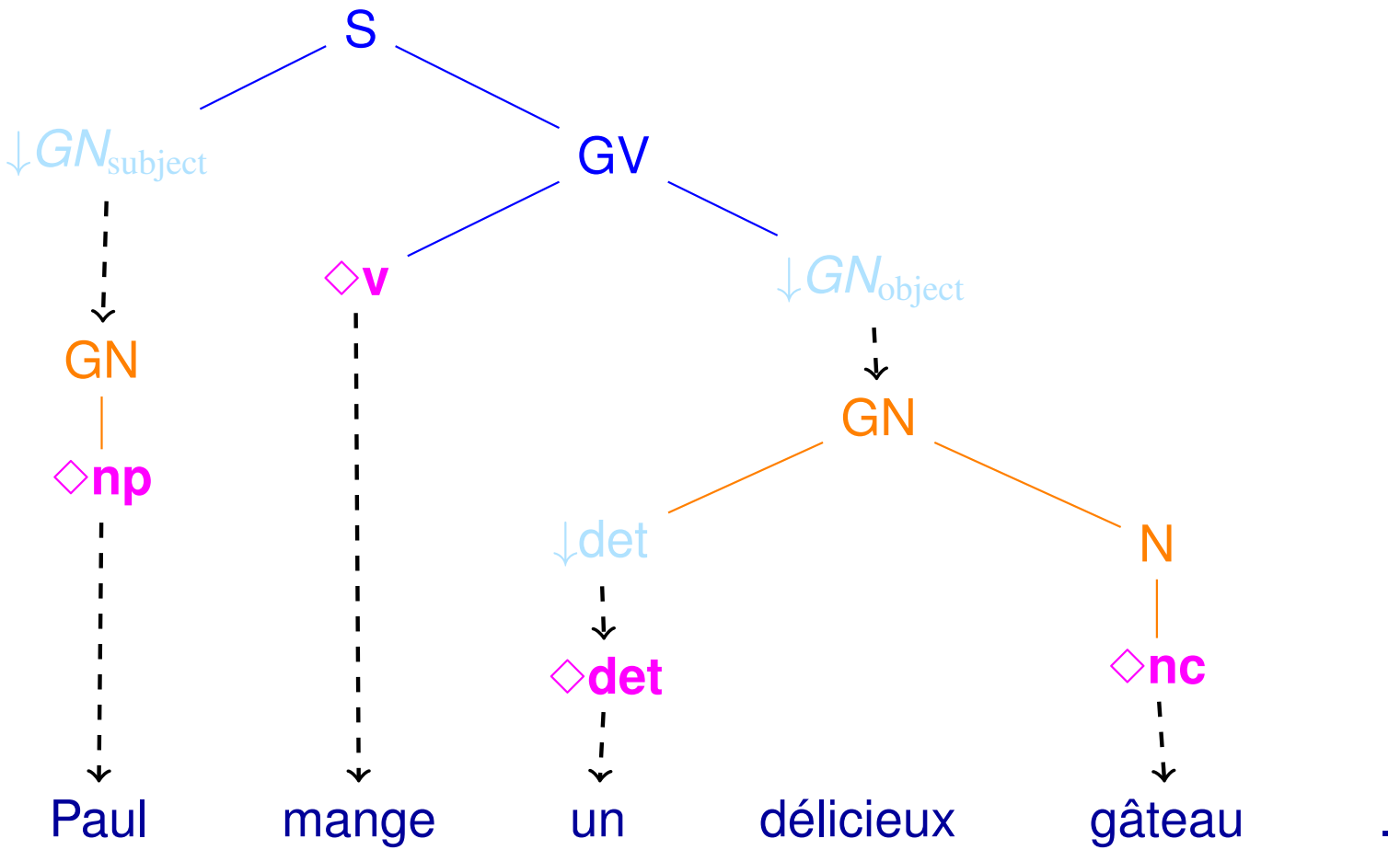
Exemple d'analyse



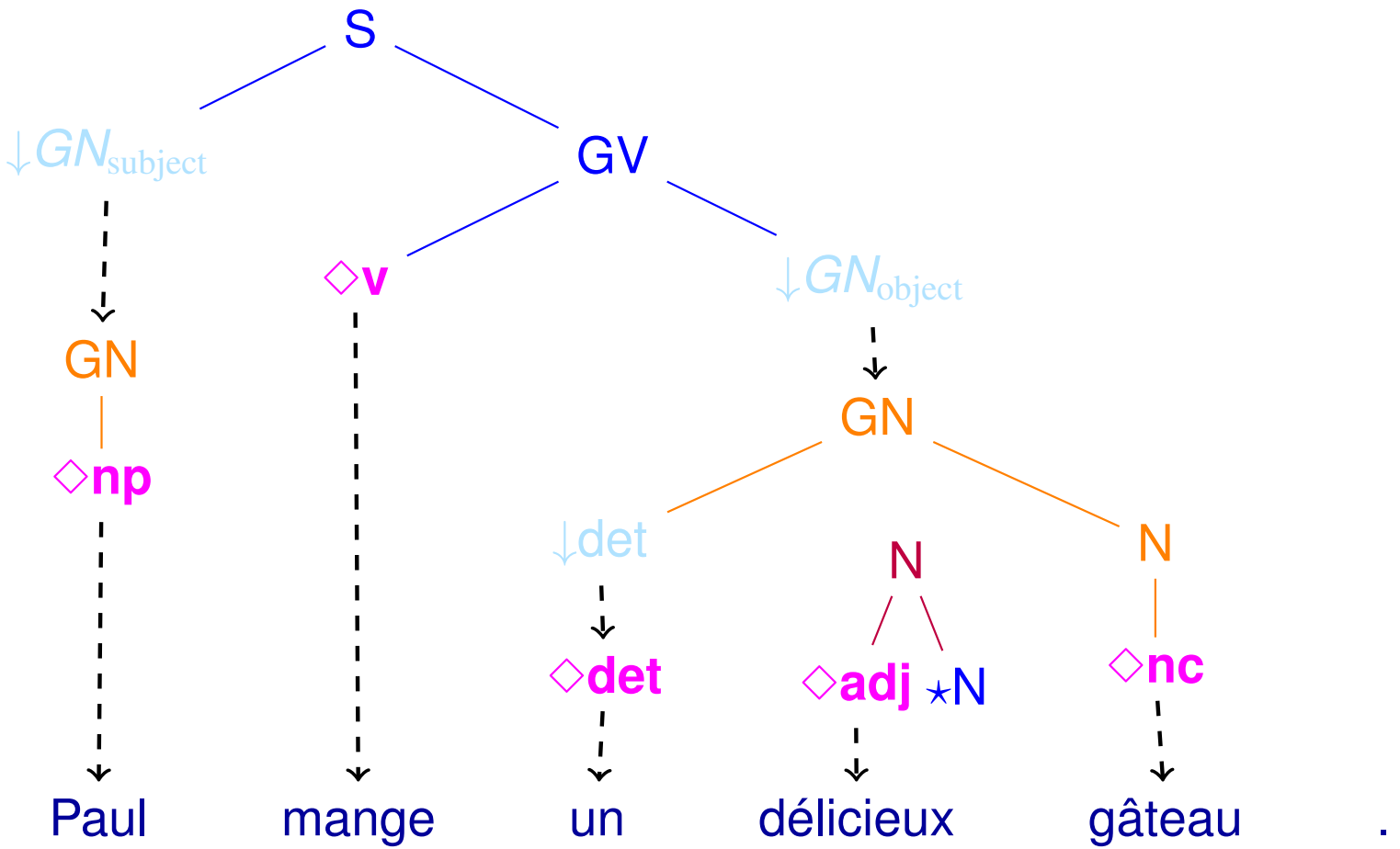
Exemple d'analyse



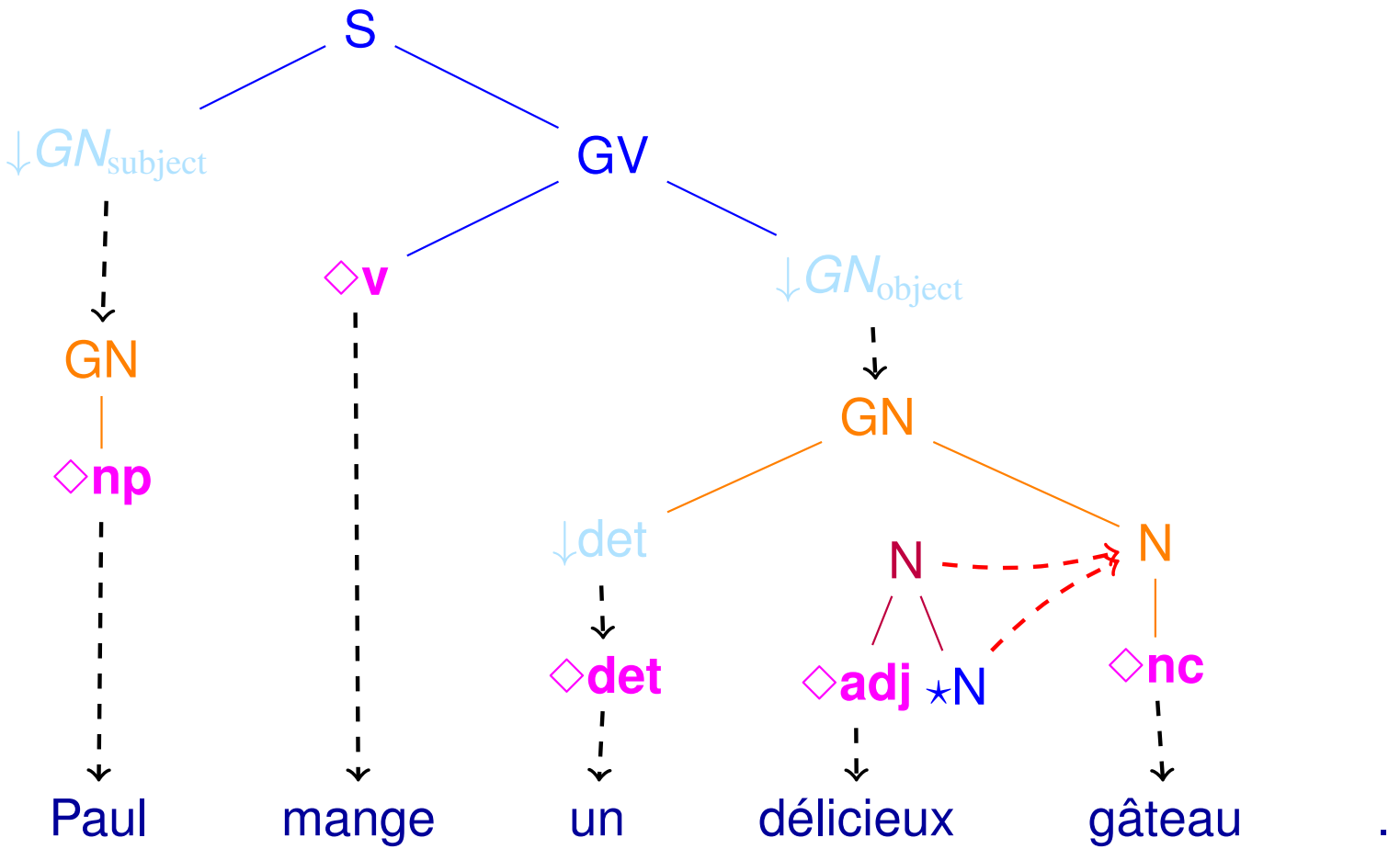
Exemple d'analyse



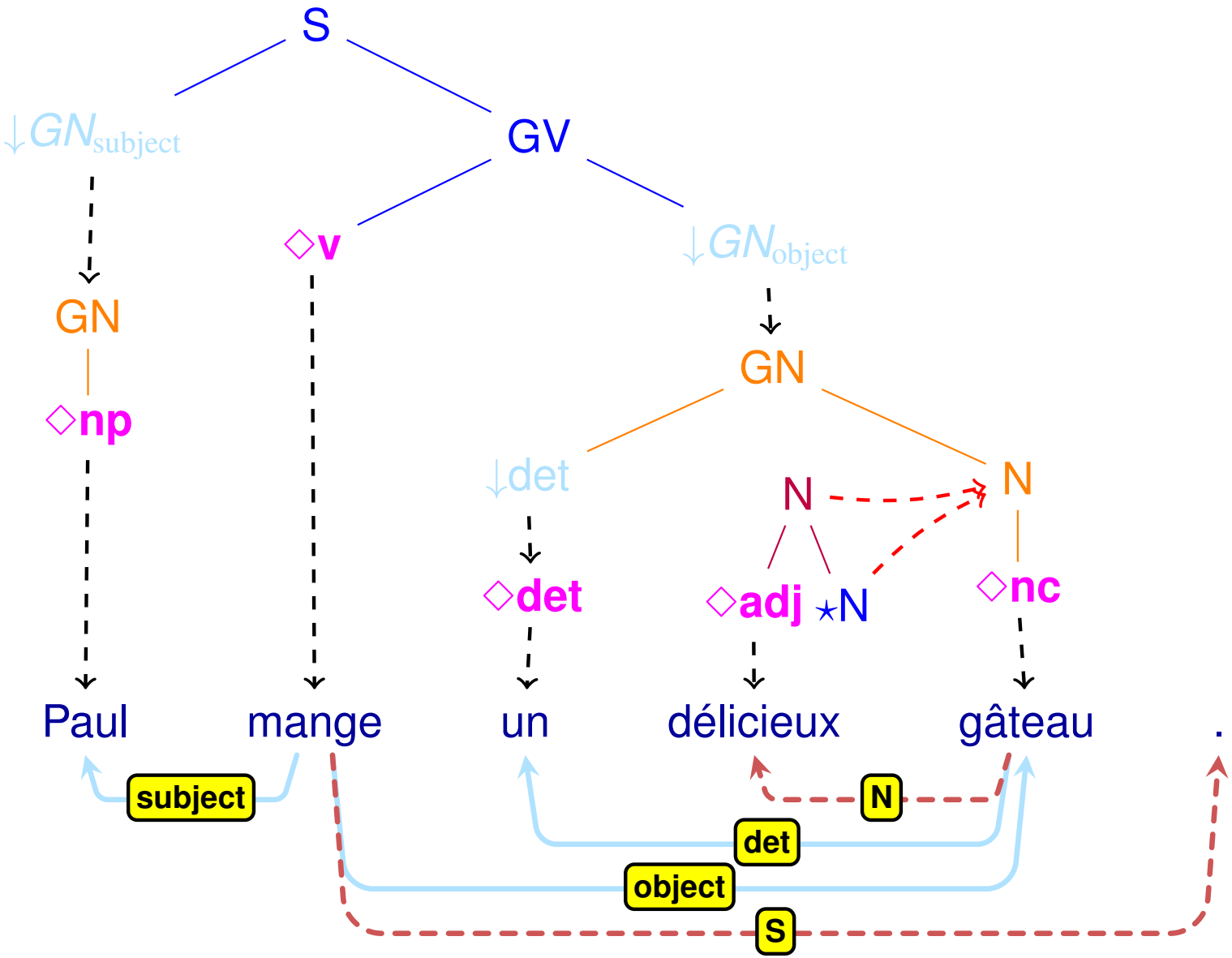
Exemple d'analyse



Exemple d'analyse



Exemple d'analyse



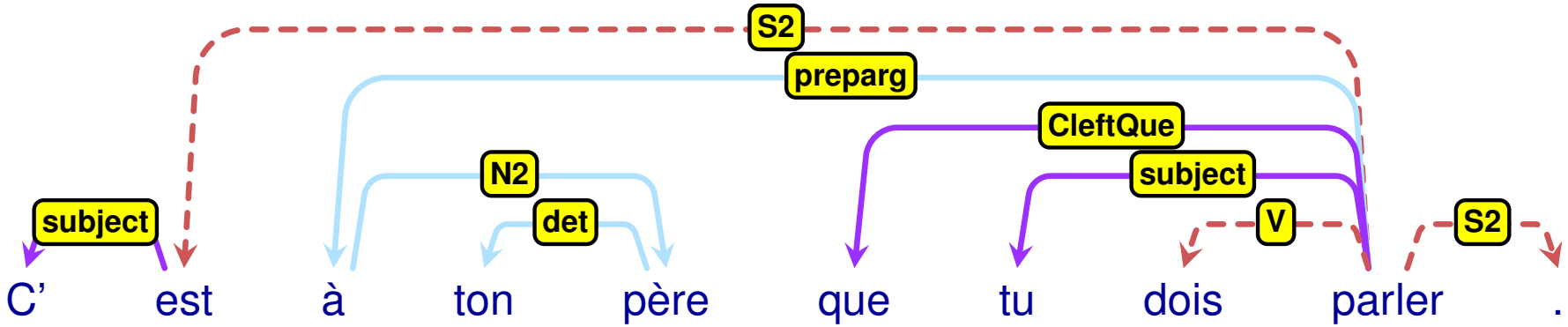
Diversité syntaxique

Les enfants allument la télé. *La télé est allumée par les enfants.*

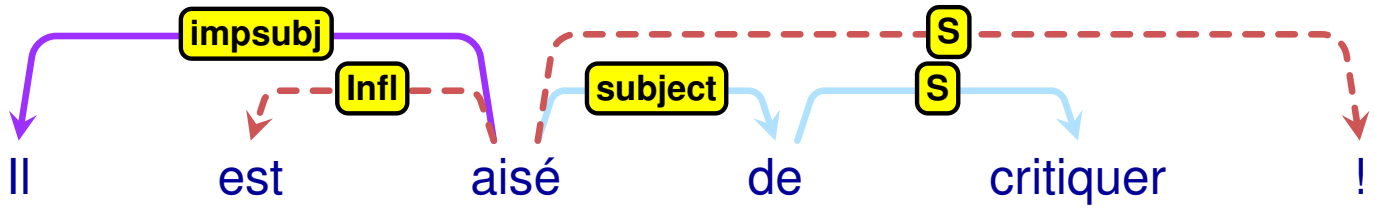
Il donne un livre à Paul. *Il donne à Paul un livre.*

Il le lui donne. *donne-le-lui ! ne le lui donne pas !*

Tu dois parler à ton père. *C'est à ton père que tu dois parler.* (*) *À ton père parler tu dois*

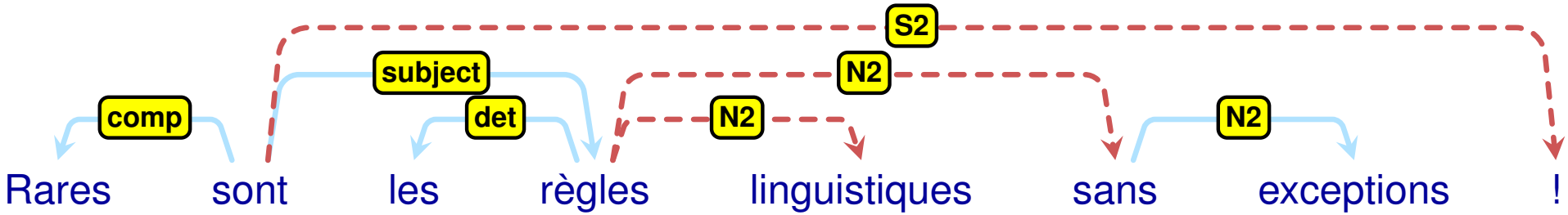


La critique est aisée. *Critiquer est aisé.* *Il est aisé de critiquer!*



Diversité syntaxique et expressions idiomatiques

Rares sont les règles linguistiques sans exceptions !



Je m'arrêtai avec, encore, l'envie de fuir.

Ne m'avouait-elle pas, il n'y a pas huit jours, que son palais l'ennuie ?

Je ne conclurai pas, exception culturelle oblige, par une citation.

Vous avez toujours la nostalgie d'un je ne sais quoi qui n'existait pas hier et qui ne sera pas demain

Qui peut le plus peut le moins

(oral) *L'état c'est moi! Le chocolat, moi, j'adore ça*

Loi de Zipf

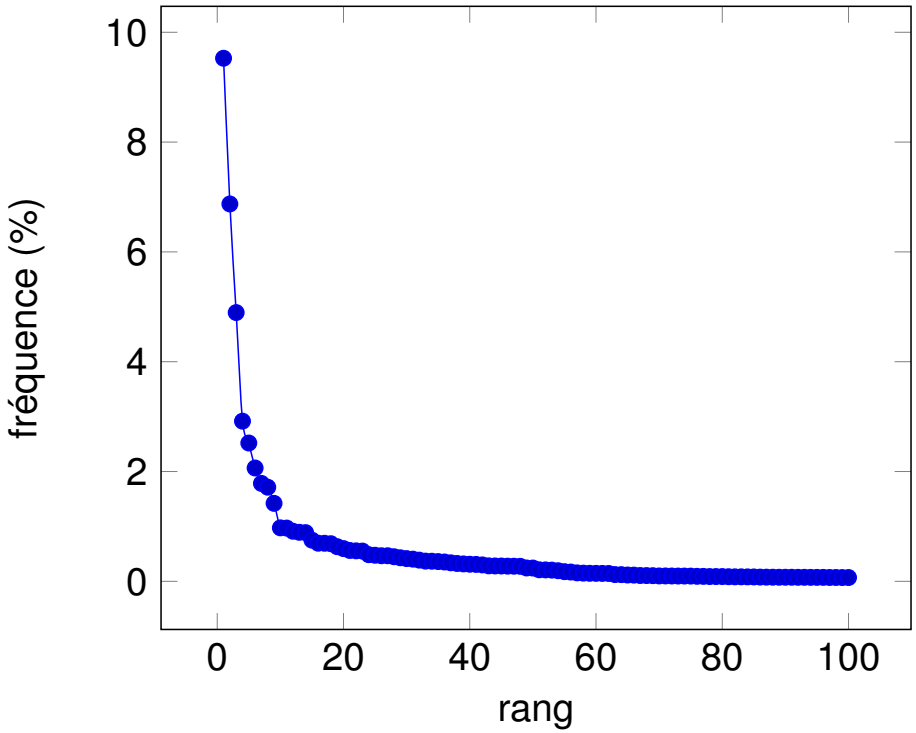
Loi en puissance très présente dans les données linguistiques, traduisant une décroissance exponentielle de la fréquence *f* en fonction du rang *n*:

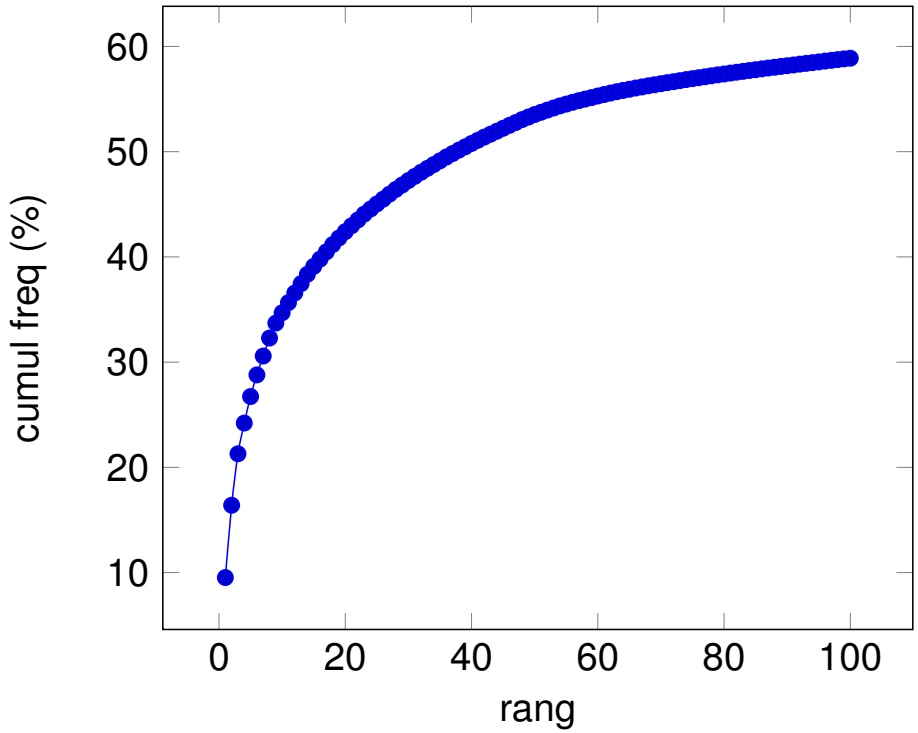
$$f_n \sim 1/n^{\alpha} \ \alpha > 1$$

- quelques mots/structures sont beaucoup utilisés; énormément de mots/structures sont très peu utilisés
- traduit à la fois une prime à la réutilisation et une tendance à la créativité

Distribution de lemmes

Distribution des mots (lemmes) dans un corpus de 500 millions de mots, avec 3 234 274 lemmes distincts dont 71 348 hors noms propres:

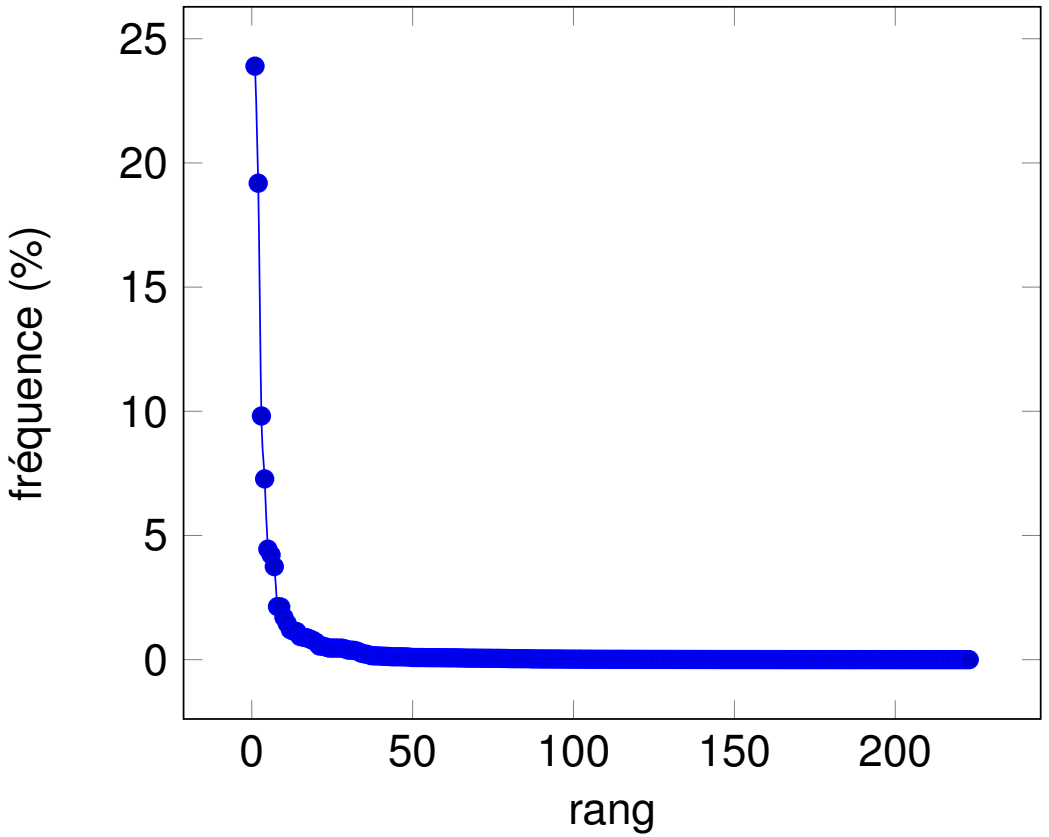




Les mots les plus fréquents: **le**, **de**, “,”, “.”, **à**, **un**, **et**, `cln`, “:”, **en**, **être**/v, . . .
80% des occurrences couvertes avec ∼1500 lemmes et 90% avec 6000 mots

Distribution des constructions syntaxiques

Distribution des arbres de la grammaire FRMG sur 10 096 phrases du corpus FRENCH TREEBANK de textes journalistiques (Le Monde).



- seulement 223 arbres sur les 344 arbres (factorisés) de FRMG sont utilisés.
- 90% des cas couverts avec 25 arbres; 99% avec 100 arbres
- FRMG a une couverture de 94.3% pour un score de 86.44% (LAS)

Les constructions les plus fréquentes: Groupes Nominaux (GN), déterminants, GP sur les noms, GP, adjectifs sur nom, GP sur verbes, construction verbale canonique, . . .

Robustesse face aux erreurs

À coté de la variabilité des constructions, il faut aussi faire face

- aux erreurs (orthographique, grammaticales, . . .)
  En particulier, la ponctuation, importante pour la structure de la phrase est souvent incorrectement utilisée: *Trois mois après cette victoire, entre le marché et les autorités monétaires les escarmouches continuent.*
- à la créativité (néologismes: **feuilletonnisation**, **s'auto-googliser**, **selfie**)
- aux entités nommées
  Autant en emporte le vent est un grand film
- aux communautés (jargons) et niveaux de langue
  t'es soin dans la swagance des bails de la douceur

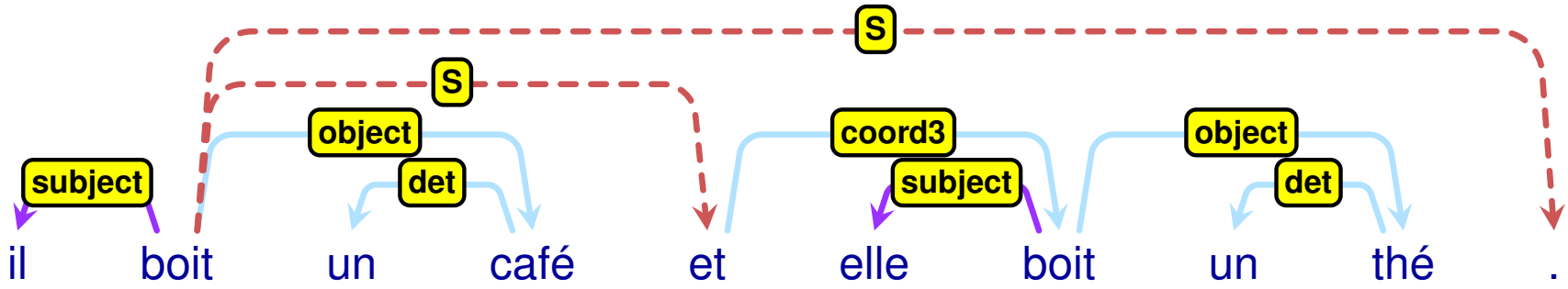
Cas de l'oral,

- pas de ponctuations, mais prosodie (dont les pauses)
- spécificités comme les hésitations (**euh**) et les répétitions
  donc moi ben je vais je je prends le mét˜ je prends le métro le matin bon jusqu'à le Palais Royal à quelle heure excusez -moi

Aux limites des grammaires

Certains phénomènes se décrivent mal avec des grammaires formelles:

- Phénomènes de structures parallèles avec des ellipses: coordination
  Il boit un café et elle un thé.



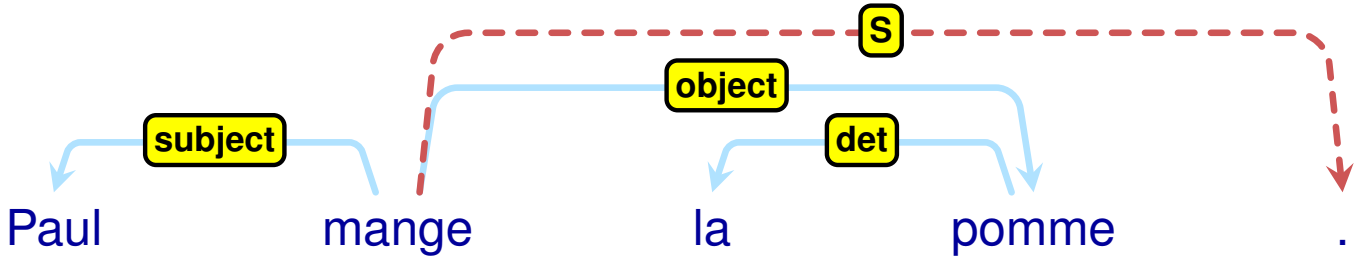
Tu dois porter les boules rouges à Paul et toi les vertes à Antoine.

- Phénomènes similaires pour l'oral avec des reprises:

```
donc < moi < " ben "
{ je vais | { je | je } prends le mét~
          | je prends le métro
} le matin
```

Ambiguïtés

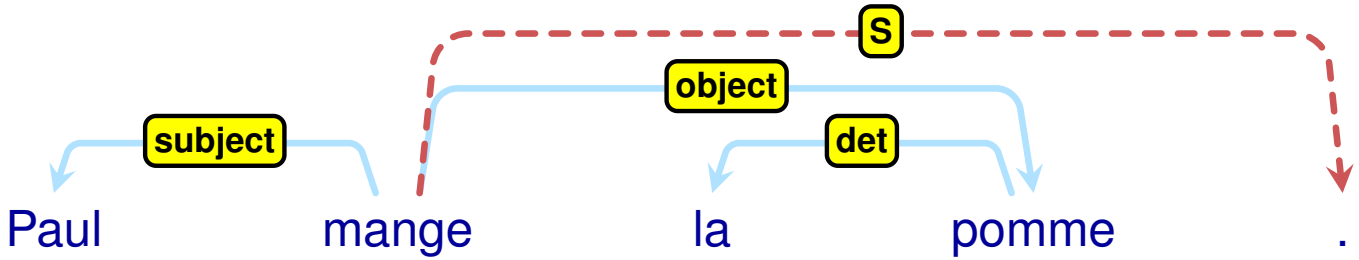
- Paul mange la pomme



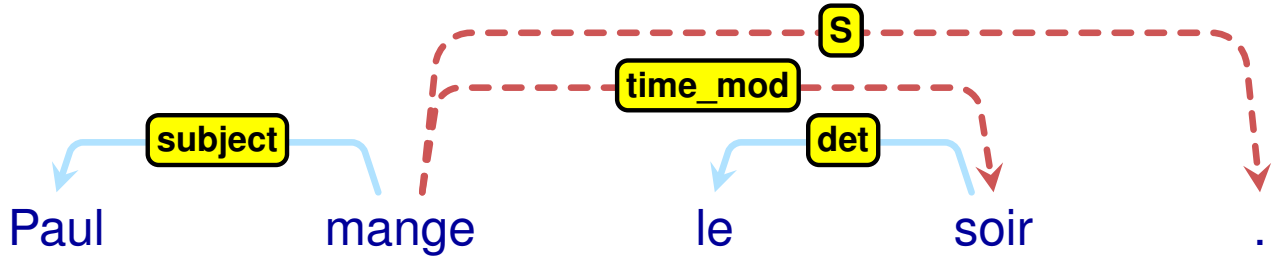
- Paul mange le soir

Ambiguïtés

- Paul mange la pomme

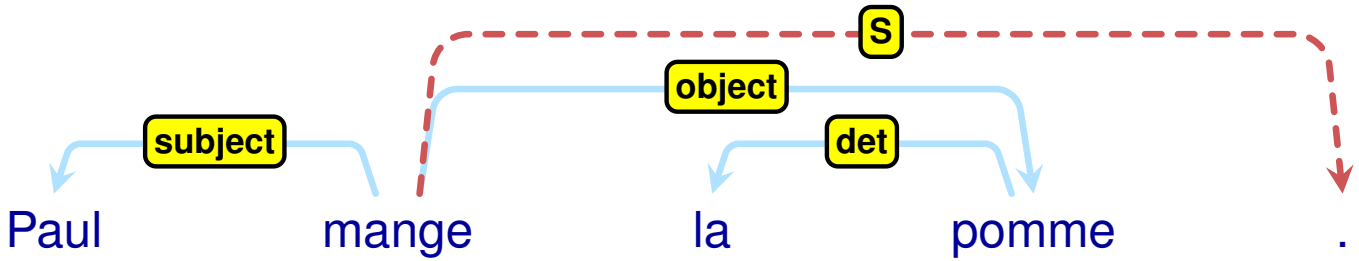


- Paul mange le soir

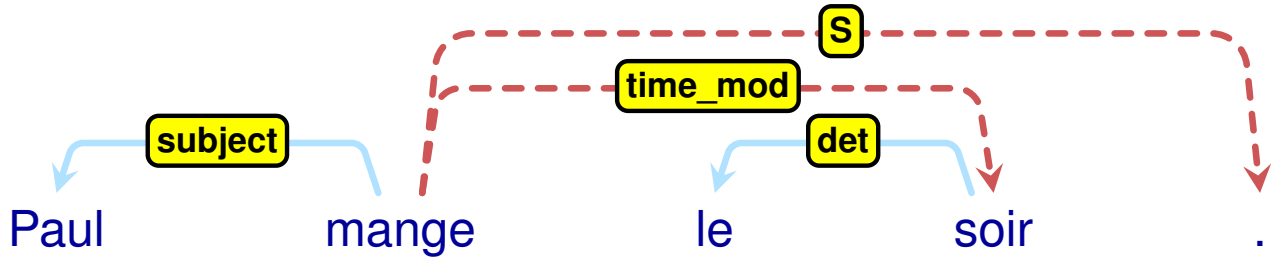


Ambiguïtés

- Paul mange la pomme



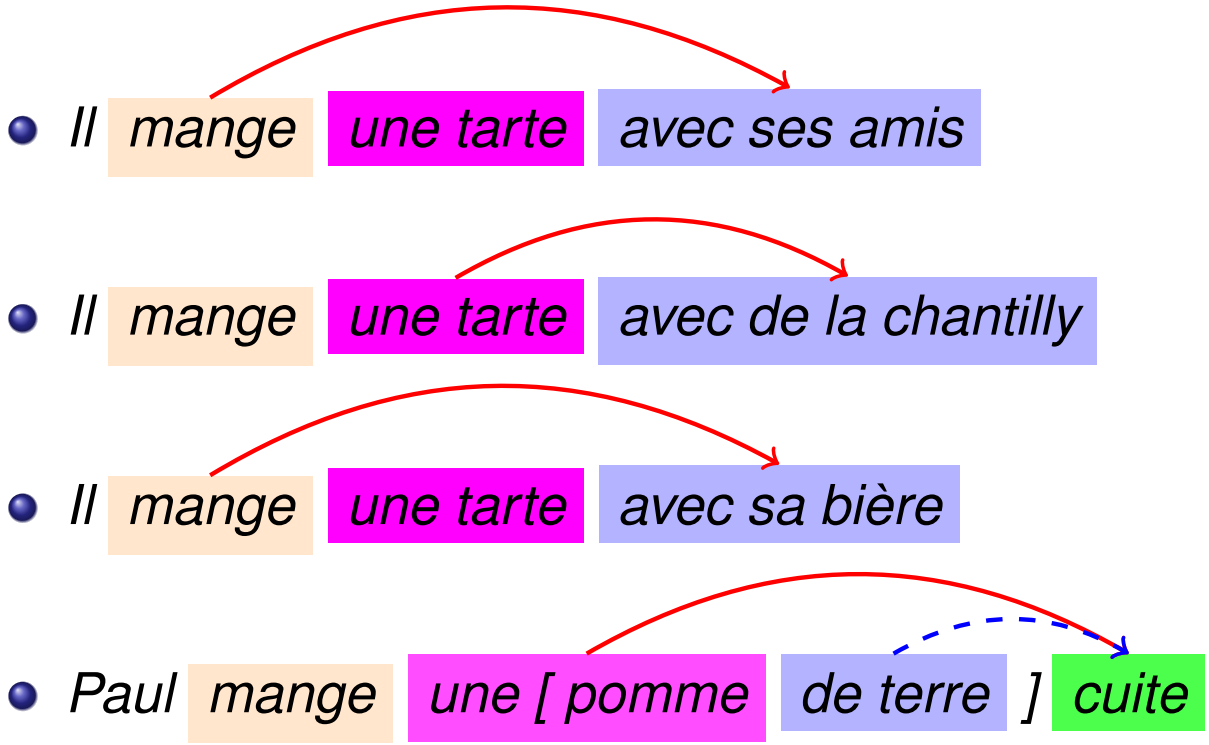
- Paul mange le soir



$\Longrightarrow$ ajout d'une construction syntaxique $\Longrightarrow$ ambiguïté
solution: ajout de traits `+time` sur **matin**, **soir**, **semaine**, ...

La malédiction des attachements prépositionnels

Ridiculement simple mais quasi-impossible à résoudre sans informations sémantiques, pragmatiques ou contextuelles:



- *Il observe une maman avec ses jumelles*

Constat: Il faut des connaissances sur l'usage des mots et des connaissances sur le monde

Stratégies de lutte

Trois grandes approches, éventuellement complémentaires, pour apporter de la connaissance:

- la construction de ressources lexicales et sémantiques fines et couvrantes
- Des approches statistiques supervisées (*treebank*)
- Des approches statistiques non-supervisées (acquisition de connaissances)

Ressources lexicales

Principe: Construire (manuellement!) des ressources linguistiques avec des informations partiellement sémantiques sur les mots.

Ainsi, FRMG utilise le lexique LEFFF (Sagot)

```
soir nc [pred="soir",@time,@ms]
```

Prototypes: *<il> promet <à quelqu'un> <de faire quelque chose>*

```
promettre v [pred="promettre<Suj:(...),
                             Obj:(de-sinf|scompl|sn),
                             Objà:(cld|à-sn)>",@CtrlSujObj,@pers,@W
```

Développement de diverses ressources plus ou moins sémantiques (WORDNET, FRAMENET, VERBNET, ...)
Mais:

- très difficile de renseigner finement pour tous les mots (+ nouveaux mots)
- bloque des usages acceptables
  j'aime le matin quand tout est encore calme.
- figures de style (métonymie, métaphores, ...)
  - *il boit un dernier verre avant de partir.*
  - *Le temps mange la vie* (Baudelaire)

Corpus annotés et apprentissage supervisé

Principe:

1. préparation de données annotées par des humains (*treebank*), éventuellement assistés par des outils.
   PENNTREEBANK, FRENCH TREEBANK, PRAGUE DEPENCENCY BANK, ...
2. apprentissage supervisé à partir des données annotées (HMM, SVN, CRF, réseaux bayésiens, ...)

Avantages:

- très efficace, autour de **90%** en analyse syntaxique
- utilisable et utilisé pour de nombreuses tâches linguistiques
  étiquetage, entités nommées, analyse syntaxique, ...
- protocole largement indépendant du langage

Désavantages:

- définition d'un guide d'annotation ($\sim$ spécification)
- coût humain important (plusieurs années.hommes)
- volumes restreints (autour d'1Mmots) et erreurs d'annotations
- pas de compréhension des grammaires acquises
- analyseur adapté au domaine d'apprentissage (souvent journalistique)

Corpus bruts et apprentissage non supervisé

Principe:

- utilisation de documents écrits par des humains (linguistiquement corrects, porteurs de sens, et reflétant les usages)
- traitement (linguistique)
- détection de redondances, corrélations, utilisation de motifs pour extraire des informations, émergence des motifs, ...

Exemples:

- utilisation à divers niveaux (apprentissage de règles morphologiques, règles de segmentation, de lexique morphologique, un peu syntaxe, ...)
- Extraction de terminologie, construction de réseau de mots, de classes de mots, acquisition de relations entre mots (`est_une_sorte_de`, `est_une_partie_de`, ...)

Du langage au sens

“Beware the Jabberwock, my son!
The jaws that bite, the claws that catch!
Beware the Jubjub bird, and shun
The frumious Bandersnatch!”

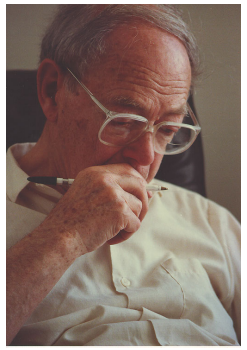
Il était grilheure; les slictueux toves
Gyraient sur l’alloinde et vriblaient:
Tout flivoreux allaient les borogoves;
Les verchons fourgus bourniflaient.

Paul s’est cassé la **binti**.
Sa fracture à la **binti** a été correctement réduite.
Il a des douleurs dans la **binti**.

Une expérience

Meanings of words are (largely) determined by their distributional patterns (Harris 1968)

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



Traitement syntaxique d'un très grand corpus et extraction des contextes:

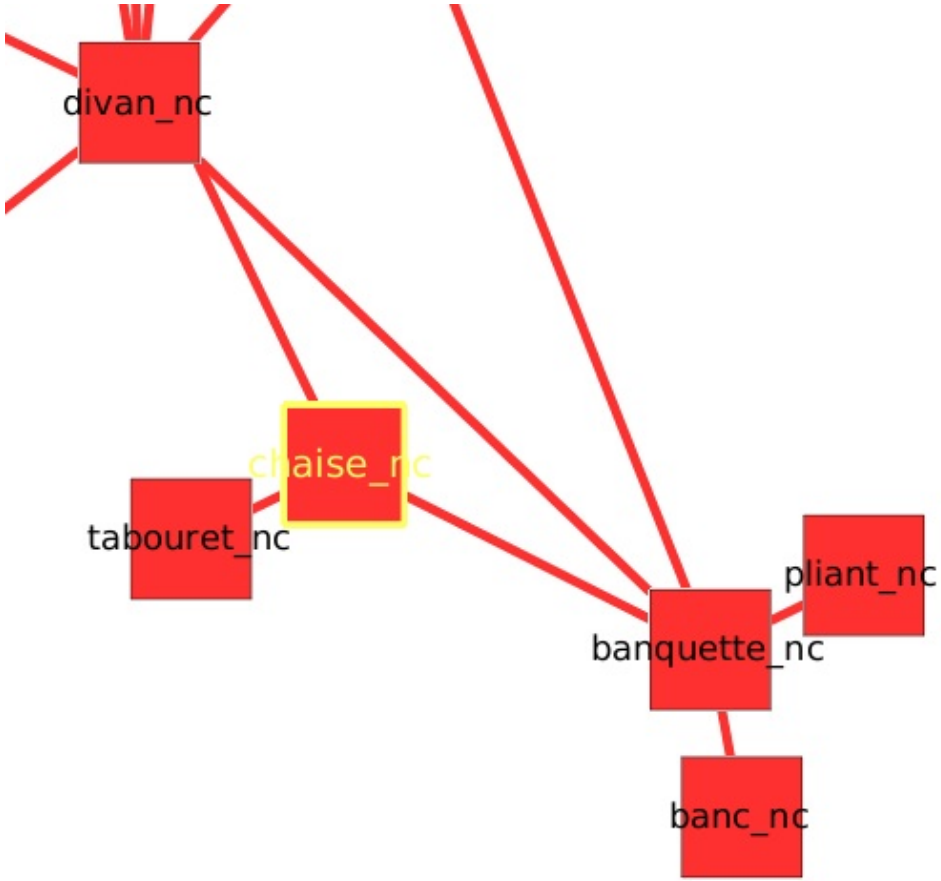
Corpus	**#phrases (millions)**	**#mots (millions)**	**Description**
Wikipedia (fr)	18.0	178.9	504K pages encyclopédiques
Wikisource (fr)	4.4	64.0	12.8K textes littéraires
EstRepublicain	10.5	144.9	journalistique
JRC	3.5	66.5	directives européennes
EP	1.6	41.5	débats parlementaires
AFP	14.0	248.3	400K dépêches
Total	52.0	744.2	

Compter et collecter des dépendances

<gouverneur>	<rel>	<régi>	<freq>
chaise_nc	et	table_nc	235
asseoir_v	sur	chaise_nc	227
chaise_nc	modifieur	long_adj	168
chaise_nc	de=	poste_nc	115
tomber_v	sur	chaise_nc	103
chaise_nc	modifieur	musical_adj	102
se_asseoir_v	sur	chaise_nc	93
prendre_v	cod	chaise_nc	87
chaise_nc	modifieur	électrique_adj	82
chaise_nc	modifieur	vide_adj	80
chaise_nc	à=	porteur_nc	80
dossier_nc	de	chaise_nc	78
avoir_v	cod	chaise_nc	71
table_nc	et	chaise_nc	62
chaise_nc	de=	paille_nc	56

Réseau de mots

Rapprochement des mots en fonction de la similarité de leurs contextes (*Markov Clustering*)



À quoi sert une chaise ?

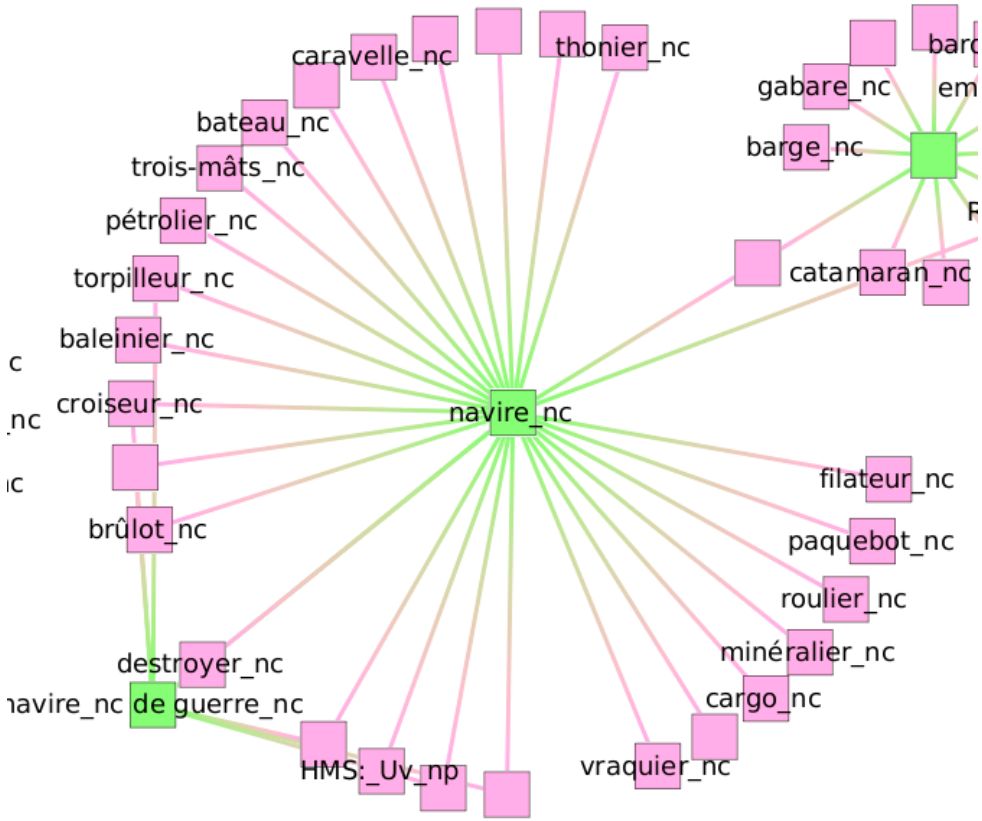
Il est possible d'examiner les contextes de rapprochement:

	chaise divan	chaise tabouret	banquette divan	banquette canapé	banquette chaise	banquette banc
se asseoir sur [•]						
asseoir sur [•]						
allonger sur [•]						
dormir sur [•]						
tomber sur [•]						
monter sur [•]						
place sur [•]						
grimper sur [•]						
installer sur [•]						
poser sur [•]						
coucher sur [•]						
siéger sur [•]						
côté sur [•]						
se lever cpl de [•]						
se affaisser sur [•]						
jeter sur [•]						
être sur [•]						
se installer sur [•]						
retomber sur [•]						
endormir sur [•]						
se soulever sur [•]						

Assigner des labels aux classes

Retrouver et analyser les définitions des mots (dans Wikipédia), avec des patrons d'extraction: *X est (une sorte de|une forme de|. . .)* Y_{genus} . . . differentia

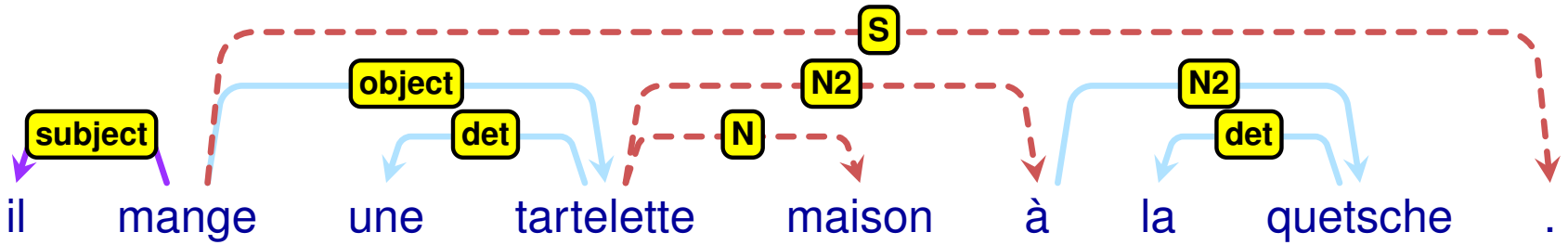
> *Un porte-avions est un navire de guerre permettant de transporter et de mettre en oeuvre des avions de combat*



Utilité

- On cherche à injecter les connaissances ainsi acquises dans les outils de traitement linguistique
  il mange une tartelette maison à la quetsche.

tartelette proche de **tarte**
quetsche sorte de **fruit**
`aux_fruits` contexte fréquent sur **tarte**
$\Longrightarrow$ **tartelette à la quetsche**



- $\Longrightarrow$ Cercle vertueux entre langage et connaissance
- Connaissances également utiles aux humains
  domaines spécialisés [jargons], recherche documentaire, recherche d'information, aide à la création d'ontologies, ...
- Mais besoin d'une phase de validation humaine

Conclusion

Le TAL commence à trouver sa place dans quelques belles réalisations et devient de plus en plus présent dans nos vies.

Le langage permet progressivement aux ordinateurs d'accéder aux connaissances et de les organiser.

Peut-être une question: les humains comprennent-ils le langage ?

- les langues sont-elles seulement des artefacts culturels ou existe-t-il une notion de grammaire universelle (Chomsky)
- notre apprentissage du langage est-il purement statistique ou existe-il des modules biologiques spécifiques ?

Je ne connais pas la moitié d'entre vous autant que je le voudrais, mais j'aime moins la moitié d'entre vous à moitié moins que vous ne le méritez !