Intégrer des ressources lexicales et grammaticales externes dans des analyseurs partiels probabilistes

Matthieu Constant* et Isabelle Tellier**

* LIGM, ** Lattice

Atelier MIXEUR 2013
21 juin 2013

Introduction

Tâches de segmentation

- Niveau lexical : repérage des mots composés et étiquetage grammatical
- Niveau syntaxique : chunking

Méthode : synthèse de différents articles

- Couplage d'analyseurs statistiques et de ressources symboliques lexicales et grammaticales
- Ressources symboliques acquises manuellement ou automatiquement
- Utilisation de techniques à états finis

Segmentation lexicale

Mots composés

- Séquences de mots non-compositionnelles sémantiquement
- Forment des unités lexicales (*table ronde* $\rightarrow$ nom, *tout à fait* $\rightarrow$ adverbe)
- Identification linguistique par tests syntaxiques et sémantiques (M. Gross 86, G. Gross 96)
- Gradation : *cordon bleu*, *vin rouge*

Identification des mots composés

- Consultation de ressources lexicales (Gross 89)
- Apprentissage supervisé (Vincze et al. 2011, Green et al. 2011)
- **Cet exposé :** combinaison des deux approches

Exemple

Analyse classique

L’	illustre	physicien	observe	un	trou	noir
D	A	N	V	D	N	A
XN			XV	XN		XA

Analyse avec reconnaissance des mots composés

L’	illustre	physicien	observe	un	trou	noir
D	A	N	V	D	N	
XN			XV	XN		

Segmentation statistique = tâche d'étiquetage

Annotation de type BIO

L'	illustre	physicien	observe	un	trou	noir
B-D	B-A	B-N	B-V	B-D	B-N	I-N
B-XN	I-XN	I-XN	B-XV	B-XN	I-XN	I-XN

Champs Markoviens Aléatoires linéaires (CRF)

- Modèle discriminant d'annotation de séquences
- Utilisation de traits : ex. suffixes pour aider l'étiquetage des mots inconnus ; catégories grammaticales prédites pour le chunking

Ressources symboliques

Dictionnaires

- Listes de formes fléchies de mots simples et composés
- Informations linguistiques associés : catégorie grammaticale, traits morphologiques et sémantiques, structure interne des mots composés, etc.

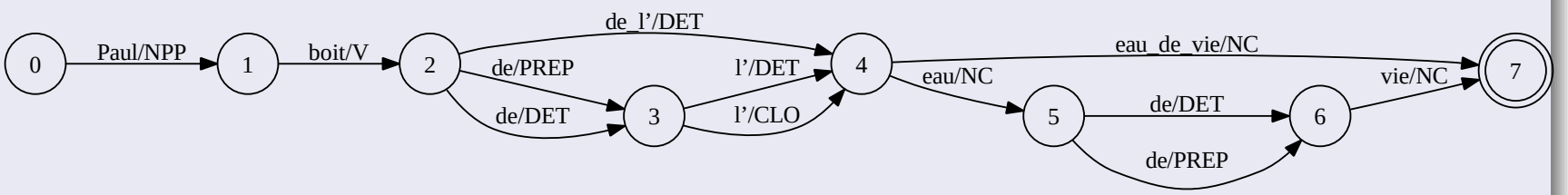
Grammaires locales/transducteurs finis

- Variations lexico-syntaxiques de mots composés
- Représentation de constituants syntaxiques simples

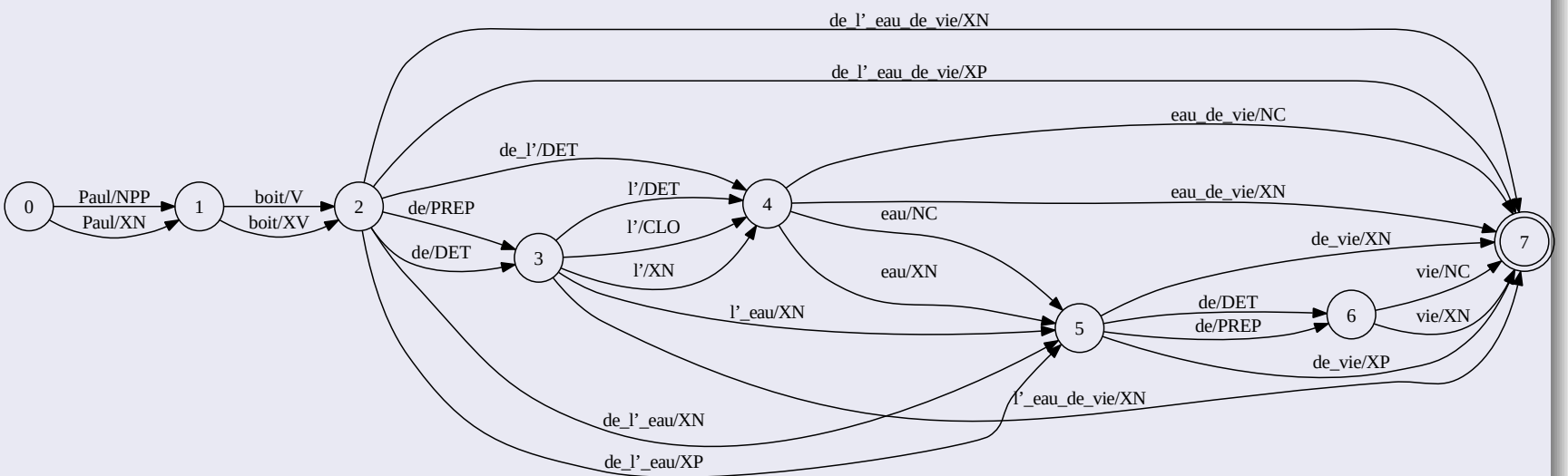
Analyse symbolique ambigue

Paul boit de l'eau de vie

Segmentation lexicale ambigue par consultation de dictionnaires



Segmentation syntaxique ambigue par cascade de transducteurs

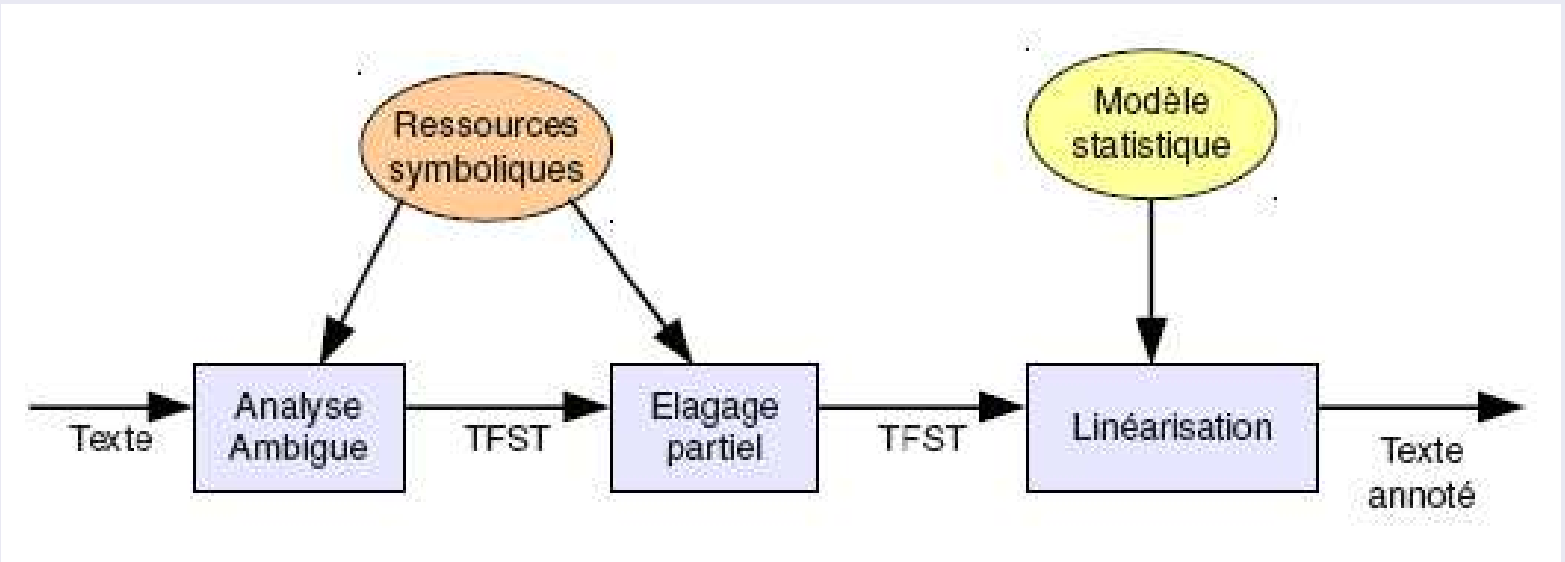


Couplage des analyseurs statistiques et des ressources symboliques

- Approche séquentielle : limitation de l'espace de recherche
- Approche combinée : source de traits du modèle CRF à la (Denis et Sagot 2009)

Approche séquentielle

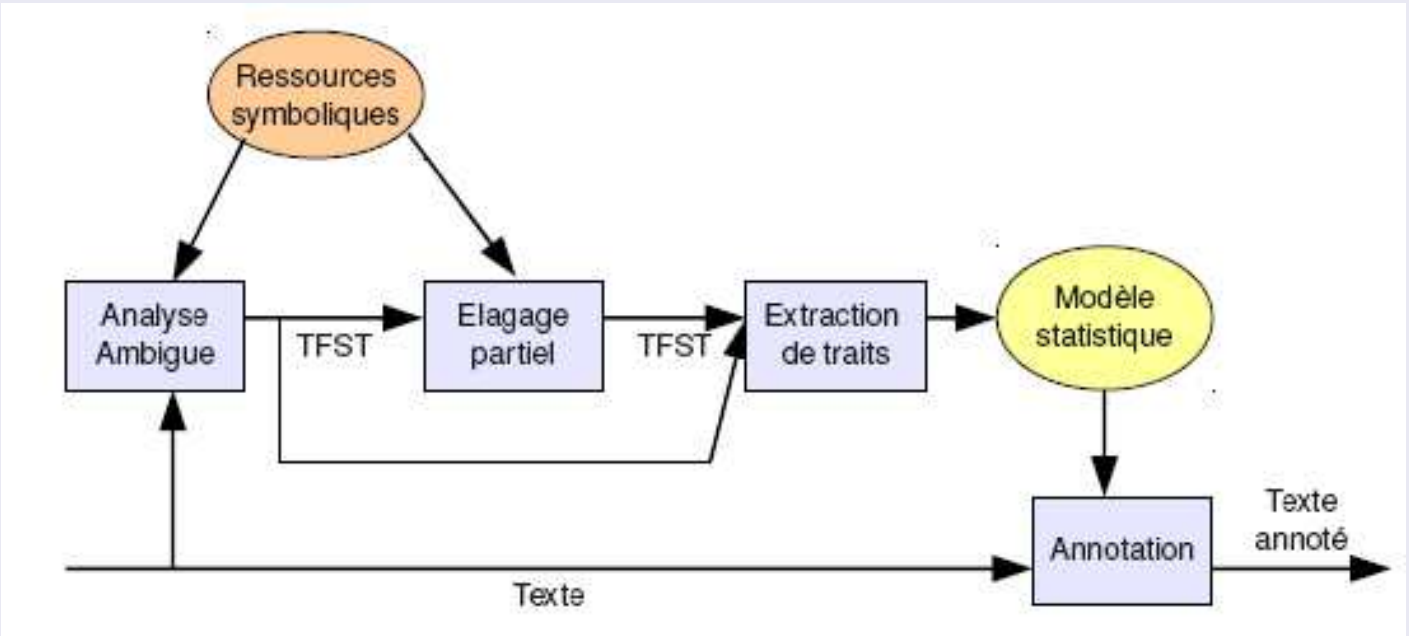
Limitation de l'espace de recherche : architecture à états-finis



- TFST = Automate représentant l'analyse ambigue du texte
- Linéarisation de l'automate par recherche du plus court chemin
- Inconvénient : ne permet pas la généralisation

Approche combinée

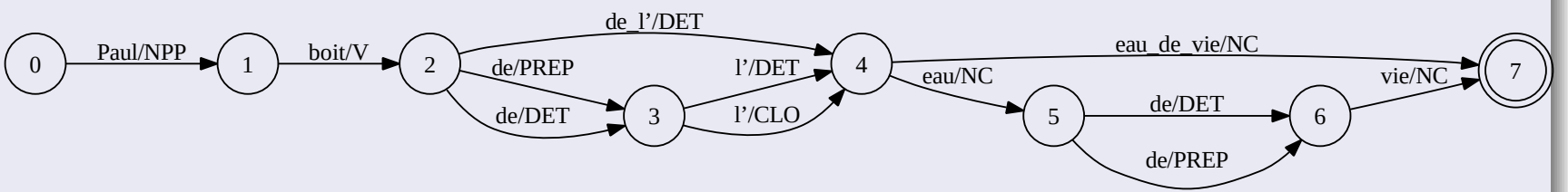
Utilisation de traits exogènes



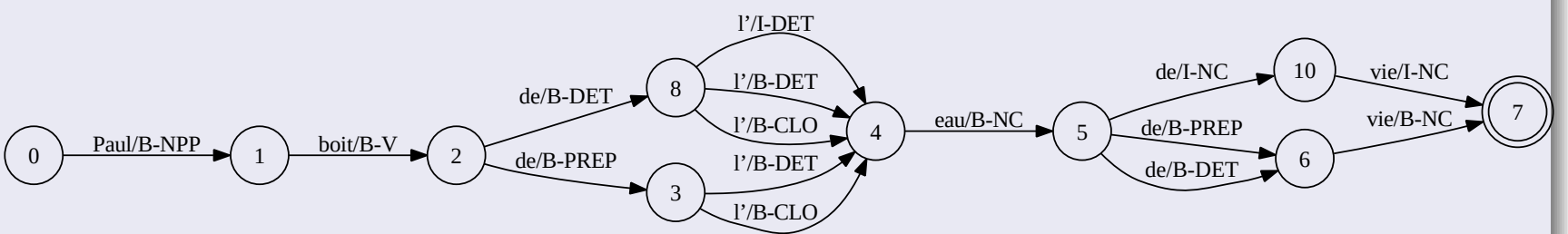
- Une position du texte $\rightarrow$ un ensemble de transitions dans TFST
- Cet ensemble permet de former soit une classe d'ambiguïté ; soit un ensemble d'attributs booléens (Constant et Tellier LREC'12)

Exemple : étiquetage grammatical

Analyse ambigue par consultation de dictionnaires

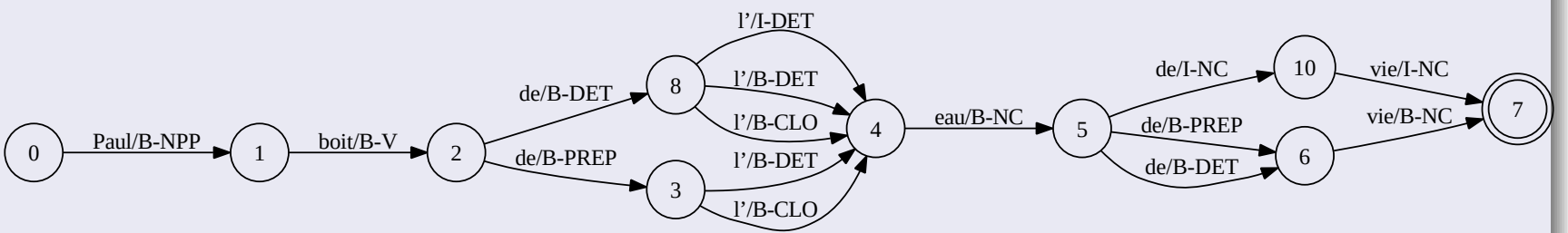


Analyse ambigue au format BIO



Exemple : étiquetage grammatical (suite)

Analyse ambigue au format BIO

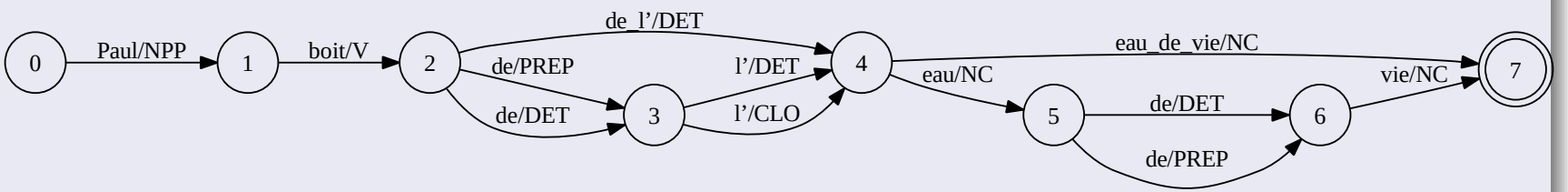


Extraction de traits

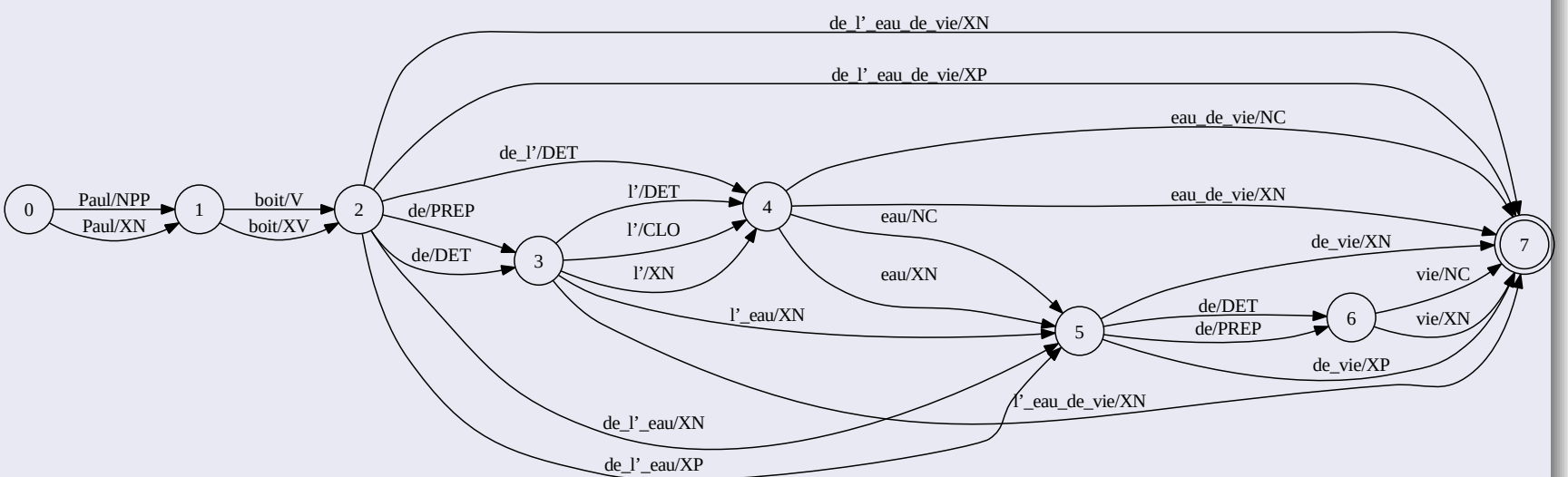
- A la position "vie", un ensemble de transitions {B-NC, I-NC}
- Equivalent à la fusion des états 6 et 10
- Attributs utilisables
  - Concaténation *B-NC_I-NC*
  - booléens : *B-NC=1* ;*I-NC=1* ; *B-V=0* ; *I-V=0* ; etc.

Exemple : segmentation syntaxique

Analyse ambigue par consultation de dictionnaires

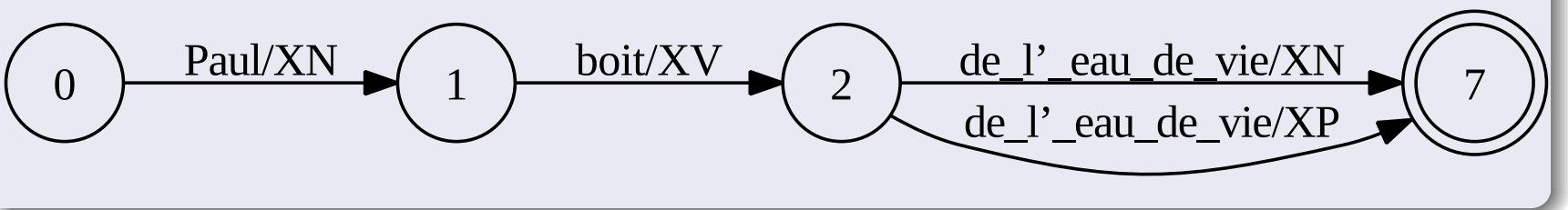


Analyse ambigue par cascade de transducteurs



Exemple : segmentation syntaxique (suite)

Elagage par plus court chemin

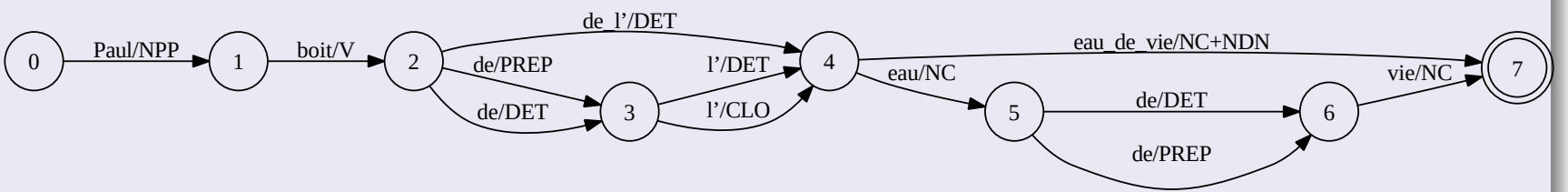


Filtrage pour l'extraction de traits

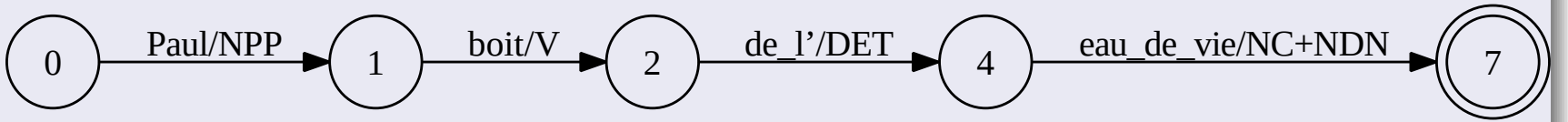
- A une position donnée, on peut filtrer les informations de différentes manières
- Filtrage des transitions
- Filtrage des informations linguistiques dans les transitions

Exemple : étiquetage grammatical

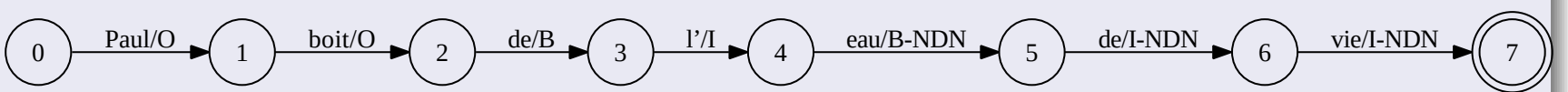
Analyse ambigue avec informations supplémentaires



Elagage par plus court chemin



Au format BIO après filtrage



Quelques chiffres

Paramètres expérimentaux

- Corpus : French Treebank (FTB)
- Approche combinée
- Extraction des traits exogènes : concaténation des étiquettes

Etiquetage morphosyntaxique

	Tous		Inconnus	
	avec	sans	avec	sans
Mots composés	78	74.5	40	31
Etiquetage grammatical	94.5	94	74.5	72

- Mesure d'évaluation = F-mesure
- avec : avec ressources lexicales
- sans : sans ressources lexicales

Bilan et perspectives

Bilan

- Approche combinée meilleure pour la segmentation que l'approche séquentielle
- Extraction des traits : pas de différence entre concaténation des étiquettes et attributs booléens
- Elagage par plus court chemin + filtrage : des petits apports

Perspectives

- Tenir compte de la structure de l'automate (pas de fusion d'états)
- Tester des élagages plus fins (règles symboliques)
- Expérimenter un élagage par méthodes statistiques "non supervisées"

MERCI !

Questions, remarques ?