# Méthodes de détection automatique d'Entités Nommées

Frédéric Béchet
LIS, Aix Marseille Université

avec des emprunts à Thierry Hamon
https://perso.limsi.fr/hamon/Teaching/P13/FDT-2016-2017/Cours/FdT-4.pdf

# Plan

- Formalisation du problème
- Difficultés
- Reconnaissance par Automates à Etats Finis
- Enlever l'ambigüité ?
  - Intégration du contexte
  - Méthodes de classification
- Reconnaissance par classification de candidats
- Evaluation
  - Taux d'erreurs, Précision, rappel, F-mesure
  - Erreurs de frontières, de type

# Formalisation du problème

- Détection et Reconnaissance d'entités nommées
  - Problème de segmentation
  - Problème de typage

```
<ENAMEX TYPE="ORG"> France-Inter <\ENAMEX> , <TIMEX TYPE=TIME> 7 heures </TIMEX>.
le journal, <ENAMEX TYPE= "PERSON"> Simon Tivolle <\ENAMEX>. bonjour!
<TIMEX TYPE=DATE> lundi 7 décembre <\TIMEX TYPE=DATE>. deux incendies
<TIMEX TYPE=TIME> cette nuit </TIMEX> en région parisienne, dans une maison de
retraite de
<ENAMEX TYPE=LOCATION> Livry-Gargan en Seine-Saint-Denis <\ENAMEX>, 7 personnes
ont péri dans les flammes.
```

# Difficultés

- Imbrications des entités

*Né à* **Paris[lieu]** *le* **21 octobre 1944[date]**, **Jean-Pierre Sauvage[personne]** *a effectué sa thèse à l'*__Université de Strasbourg[lieu][organisation,lieu]__ *sous la direction de* **Jean-Marie Lehn[personne]**. *Après un post-doctorat à* **Oxford[organisation,lieu]**, *il revient en* **France[lieu]** *et effectue sa carrière au* **CNRS[organisation]** *qu'il intègre en* **1971[date]** *et devient directeur de recherche au* **CNRS[organisation]** *en* **1979[date]**. **Jean-Pierre Sauvage[personne]** *travaille à l'*__Institut de science et d'ingénierie supramoléculaire[organisation]__ *(*__CNRS[organisation]__*/*__Université de Strasbourg[lieu][organisation,lieu]__*). Il a également reçu la médaille de bronze en* **1978[date]** *et celle d'argent du* **CNRS[organisation]** *en* **1988[date]**.

# Difficultés

- La portée des classes : Clint Eastwood, l'épouse Chirac, les frères Cohen, les démocrates, les Boeings, Bison futé
- La coordination : Barack et Michelle Obama, M. et Mme Obama
- L'imbrication : Université de Strasbourg
- Les frontières : l'équipe de Nantes, le Palais Bourbon, monsieur Hollande/le président Hollande, le couple Obama
- Les variantes : l'équipe de Nantes/le stade nantais/les canaris/les nantais/Nantes/FCN
- La polysémie : Clint Eastwood (acteur, réalisateur, producteur, mais aussi chanteur jamaïquain, chanson, personne de film), Leclerc (maréchal, homme d'affaire, Char, supermarché)

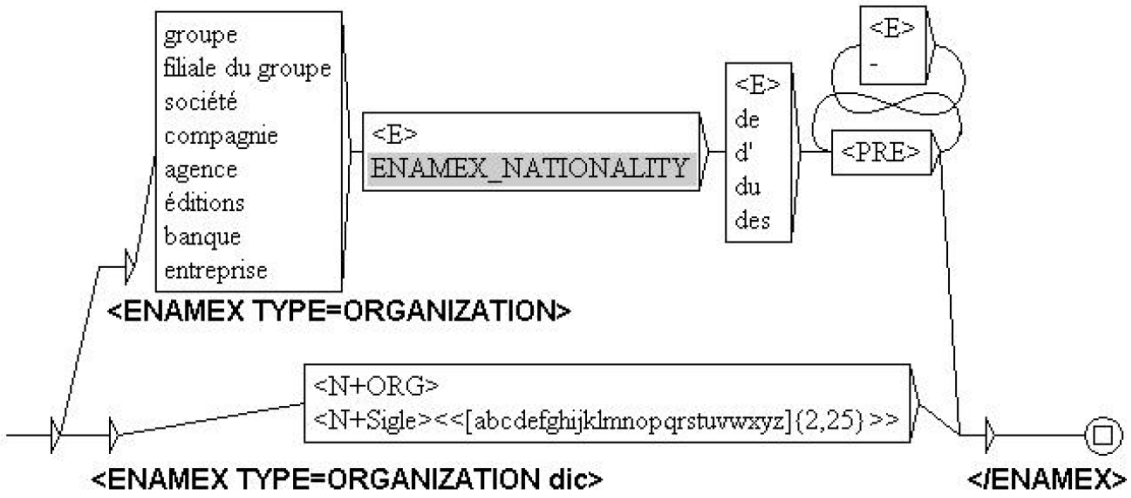
# Quelles méthodes pour détecter les entités nommés ?

- Approches à base d'étiquetage de séquences
  - N'importe quel type d'étiqueteur !!
    - CRF, RNN, modèles de bout en bout, …

| | | |
|---|---|---|
| bonjour | nc | O |
| . | ponctw | O |
| investiture | nc | O |
| aujourd'hui | adv | B-TIME |
| à | prep | O |
| Bamako | np | B-LOC |
| , | ponctw | O |
| Mali | np | B-LOC |
| , | ponctw | O |
| du | prep | O |
| président | nc | B-FONC |
| Amadou | np | B-PERS |
| Toumani | np | I-PERS |
| Touré | np | I-PERS |
| , | ponctw | O |
| réélu | v | O |
| en | prep | B-TIME |
| avril | nc | I-TIME |
| dernier | adj | I-TIME |

# Quelles méthodes pour détecter les entités nommés ?

- Approches à base d'automates
  - Représentation linguistique des entités
    - Sous forme de grammaires
    - Le plus souvent des grammaires régulières
      - Automates
      - Transducteurs



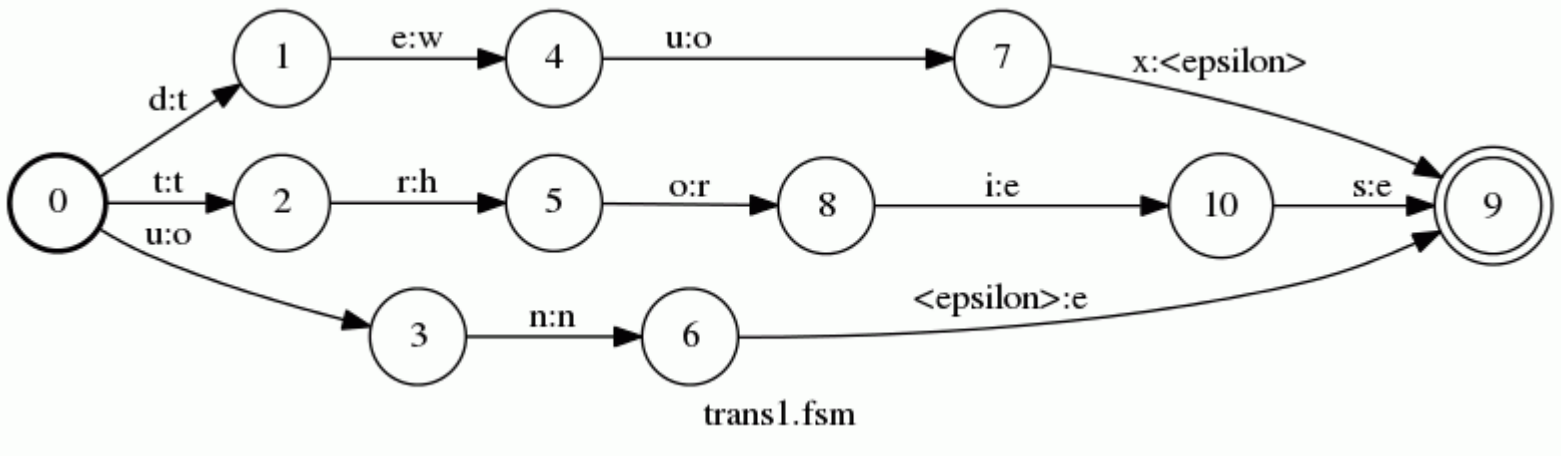

Exemple de transducteur décrivant les noms d'entreprise avec Unitex (Tolone, 2006)

# Reconnaissance par Automates à Etats Finis

- Projection de dictionnaires
  - On retrouve les entités nommées connues
  - Catégorisation des entités nommées
- Utilisation des majuscules
  *Alan Turing*, *Metro Goldwin Mayer*, *Nobody Can Beat the Wiz*
  - Indice insufisant : le premier mot des phrases est généralement en majuscule...
  - Problème de la limite à droite
    `Institut national de recherche en informatique et en automatique`
    *Organisation des Nations Unies efficace*
  - Solution : utilisation de grammaires des EN et du lexique

# Reconnaissance par Automates à Etats Finis

- Utilisation d'automates de type Transducteurs
- Avec éventuellement des poids sur les transitions



# Reconnaissance par Automates à Etats Finis

- Automates basées uniquement sur la forme des entités nommées

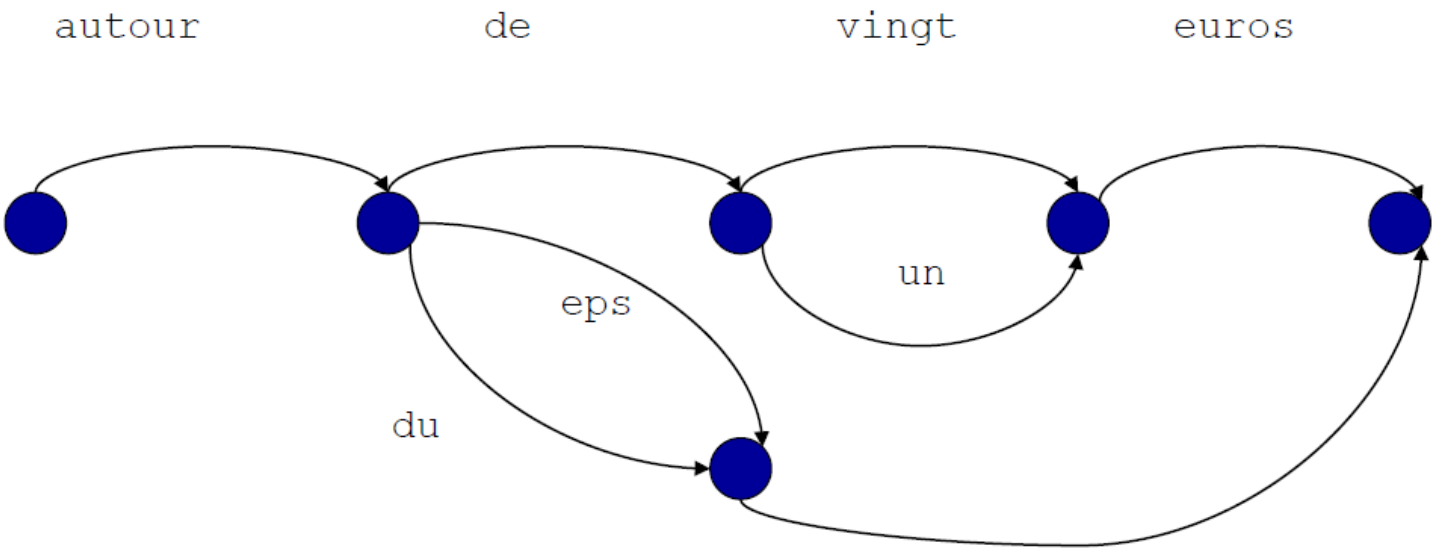
  - Utilisation des caractéristiques de la séquence
    Les entités ont une structure interne :
    - **Luc** Besson, **F.** Hollande, **H.** Clinton
    - **docteur** Jean Dupond, **maître** Durant, **président** Obama
    - Sherwood **Forest**, Hollywood **Boulevard**, **Place** de l'étoile, **aéroport** d'Orly
    - **groupe** Vivendi, **société** Général, Airbus **group**
  - Utilisation d'indices internes à l'entité
    - Majuscule, prénoms, abréviation de prénoms
    - Mots classifiant des métiers des lieux, des organisations
    - ...

# Reconnaissance par Automates à Etats Finis

- Automates intégrant des indices liés au contexte d'apparition de l'entité
  - Hypothèse : existence d'un contexte facilitant l'identification d'entités nommées et leur catégorisation
  - Utilisation du contexte locaux des entités :
    - Personne : titre, métier, grade, ...
      **juge** van Ruymbeke, **docteur** Freud, **monsieur** Chirac, **général** De Gaulle
    - Organisation : statut, activité, ..
      la **filiale** de PSA, la **compagnie** Ryanair, le **motoriste** Safran, **constructeur aéronautique** Airbus
    - Lieux :
      la **ville** de Rennes, le **fleuve** amazone, la **comète** Tchouri, le **sud** de Paris, **basé à** Lyon, **lac** Baïcal
    - Mais aussi contexte spécifique :
      Transcription of the cotB, cotC, and cotX **genes**
      la **sonde** Rosetta, le robot **Philae**

# Reconnaissance par Automates à Etats Finis

- Exemple

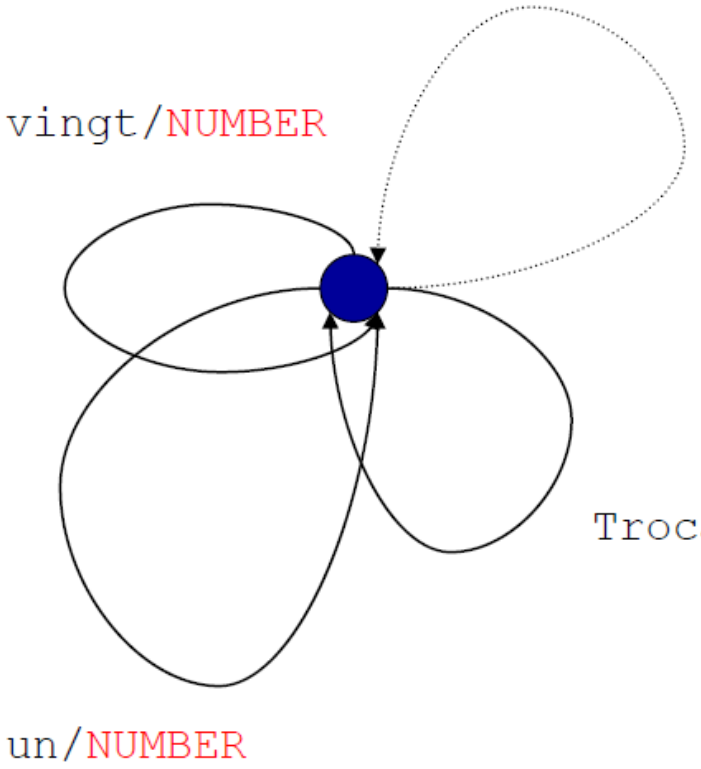



Word graph G1

Trocadéro

# Reconnaissance par Automates à Etats Finis

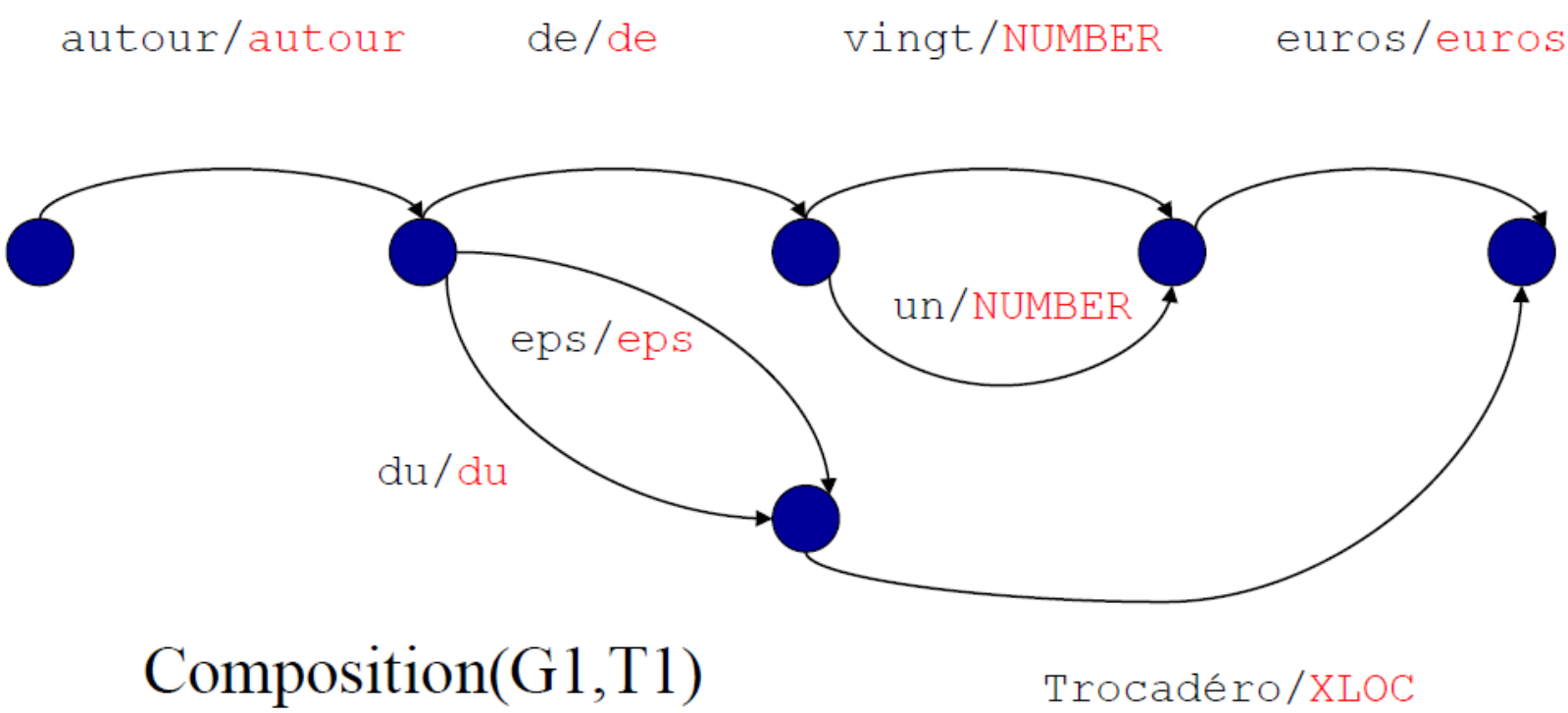
- Exemple




Generalization transducer T1

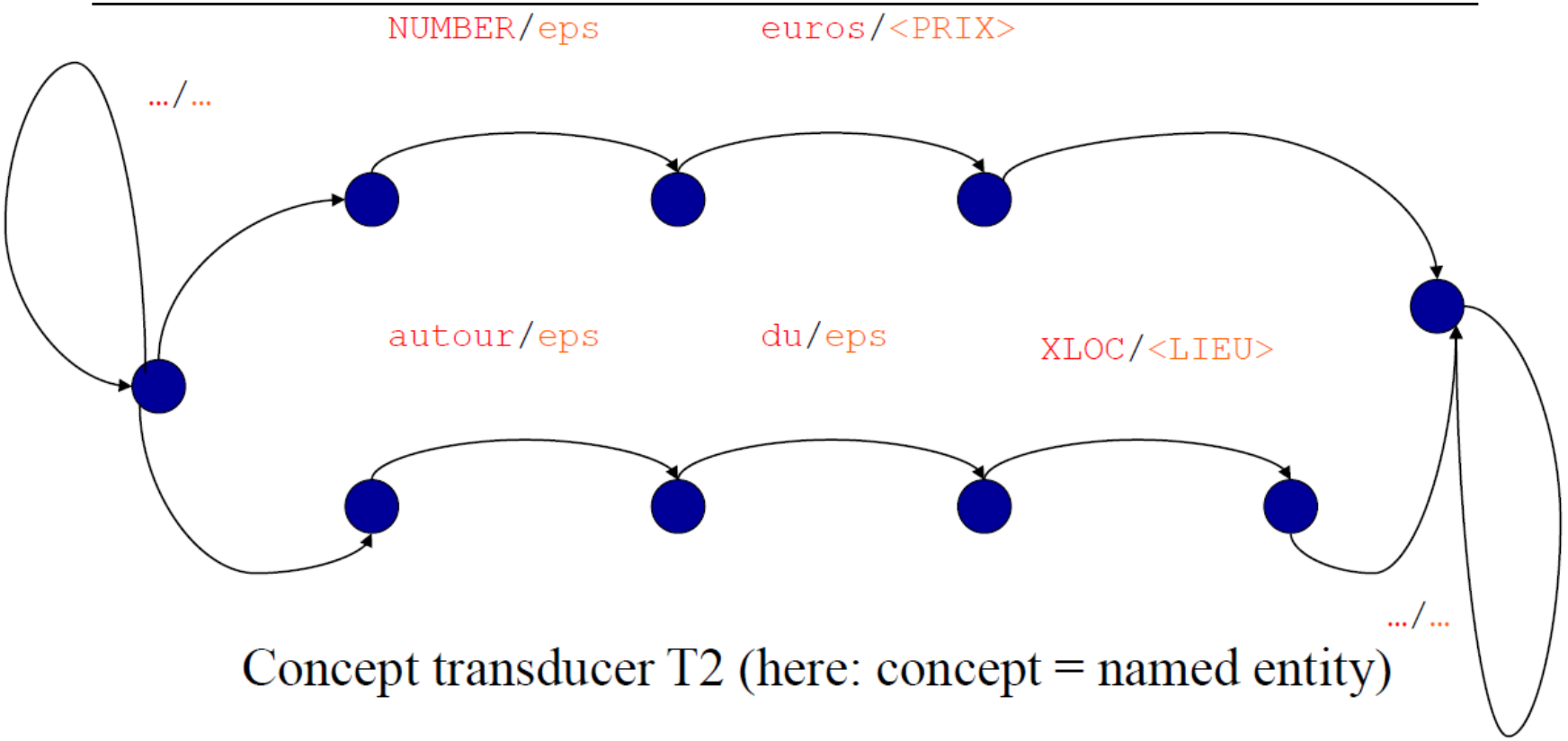
# Reconnaissance par Automates à Etats Finis

- Exemple



# Reconnaissance par Automates à Etats Finis

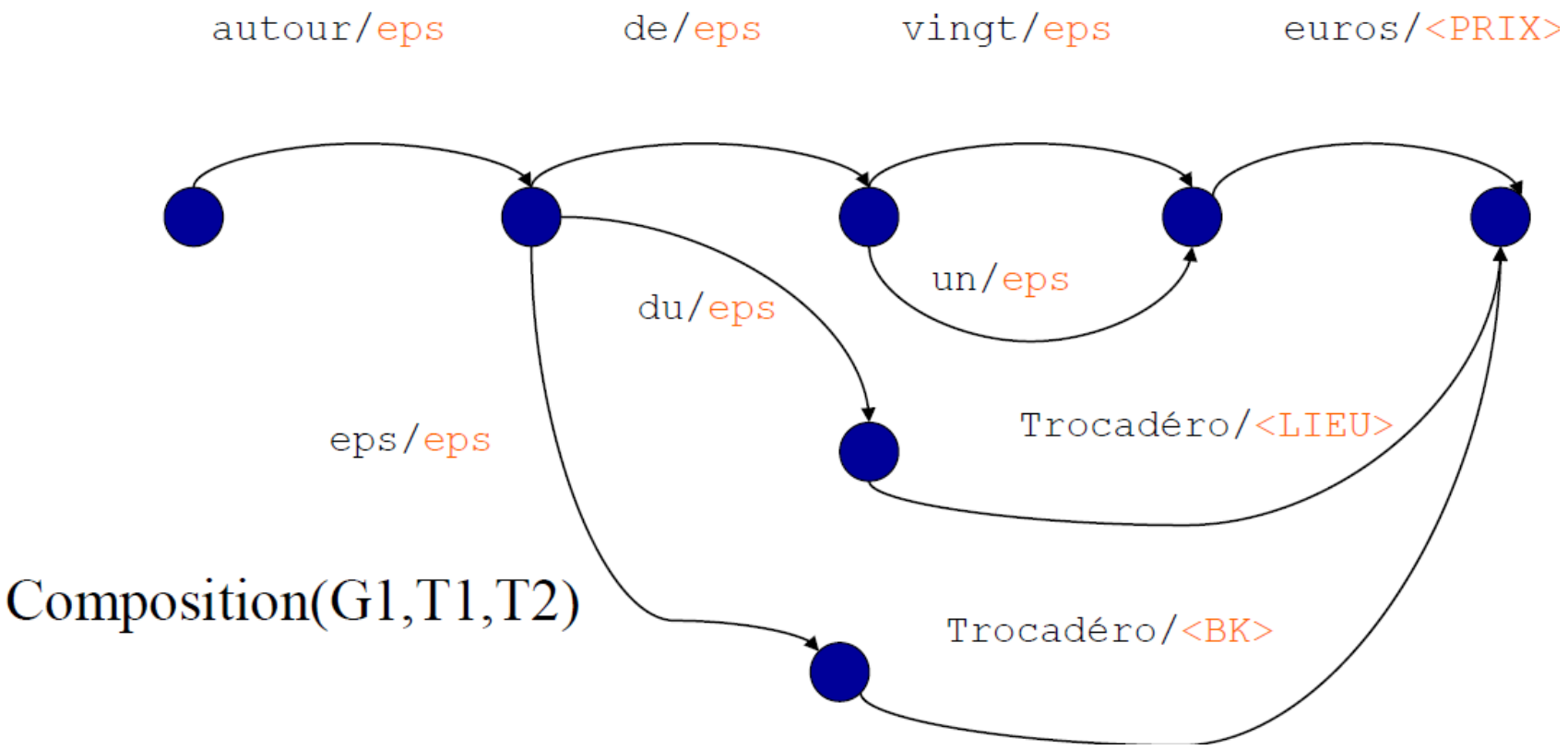
- Exemple




Concept transducer T2 (here: concept = named entity)

# Reconnaissance par Automates à Etats Finis

- Exemple



# Reconnaissance par Automates à Etats Finis

- Bilan
  - Méthode demandant beaucoup d'expertise manuelle
  - Mais se basant sur des ressources linguistiquement valides
- Avantages
  - Fiabilité des annotations
  - Possibilité de comprendre le modèle pour un humain, et de le corriger !!
- Inconvénients
  - Manque de couverture de tous les cas possible => généralisation réduite
  - Problème de l'ambigüité
    - Si on augmente le contexte, l'ambigüité diminue mais la couverture aussi !!

# Enlever l'ambigüité ?

- Intégration du contexte dans la procédure d'analyse
  - Méthodes à base de règles => problème de l'équilibre contexte/couverture

- Méthodes de classification
  - Se placer dans un cadre d'apprentissage automatique
  - Détection de « candidats entités nommées »
  - Classification en type d'entités ou bien rejet du candidat
  - Quelles méthodes de classification ?
    - Grand choix en magasin !!!!
    - Arbre de décision, Support Vector Machine, perceptron, Boosting, ….
    - Nous allons utiliser le boosting en TP : ICSIBOOST https://github.com/benob/icsiboost

# Reconnaissance par classification de candidats

- Formatage de corpus annotés « par colonne »
  - 1 mot (token) par ligne
  - chaque colonne décrit le token
- Exemple

```
0       investiture     21599   N       O
1       aujourd'hui     3625    ADV     O
2       à               21      PREP    O
3       Bamako          4288    NP      B-geoloc
4       ,               1       YPFAI   O
5       Mali            24448   NP      B-geoloc
6       ,               1       YPFAI   O
7       du              13167   PREP    O
8       président       31135   N       O
9       Amadou          1783    NP      B-person
10      Toumani         39548   NP      I-person
11      Touré           39557   NP      I-person
```

# Reconnaissance par classification de candidats

- Détection de candidats entités nommées
  - Par des automates sur les mots
  - Par des « patrons » formés de mots + catégories
    - ex: NUMBER, PRENOM, VILLE, PAYS, NOM-PROPRE, etc.
  - **Par des « patrons » de catégories morphosyntaxiques**
- Exemple de patrons
  - président Amadou Toumani Touré => N NP NP NP
  - association des producteurs de pétrole africains => N PREP N PREP N A
- Mais aussi
  - N PREP N PREP N A => situation des vendeurs de légumes verts
- Donc ambigüités !!
  - 1 patron peut déclencher un groupe nominal qui n'est pas une entité nommée
  - Trouver ensuite la bonne étiquette en fonction du contexte d'occurence

# Evaluation

- A suivre au prochain épisode !!