LA CONTRIBUTION DES GRAMMAIRES ET DU PARSING À LA RECHERCHE D'INFORMATION

Alexis Langlois – IFT6010

Sommaire

- Candidats pertinents des grammaires
  - Qu'est-ce qu'une grammaire générative/d'unification?
  - Parseurs
- Processus de recherche et intégration des candidats
  - Indexation et pondération
  - Expansion de requêtes
- Contribution
  - Travaux
  - Alternatives?

Candidats pertinents des grammaires

Grammaires génératives

Grammaires génératives

- LFG (Lexical Functional Grammar)
  - (Kaplan et al., 1994) - (Dalrymple, 2006)
    - Features des structures fonctionnelles
- HPSG (Head-driven Phrase Structure Grammar)
  - (Pollard et Sag, 1994) – (Levine et Meurers, 2006)
    - Features de tête lexicalisée
- Autres modèles
  - (Collins, 1997)

Grammaires génératives

Qu'est-ce qu'une grammaire générative?

- Une grammaire hors-contexte faisant intervenir des traits syntaxiques et lexicales dans ses règles (e.g. accord/genre).
- Formellement, ces traits sont vues comme des features au sein d'un modèle.
- Résultats: en plus d'intervenir dans la composition syntaxique, les non-terminaux deviennent une source d'information pour un amalgame de tâches d'extraction.

Grammaires génératives - LFG

- Le formalisme des LFG est établi selon deux niveaux de représentation:
  1. **Structure des constituants (c-structure)**

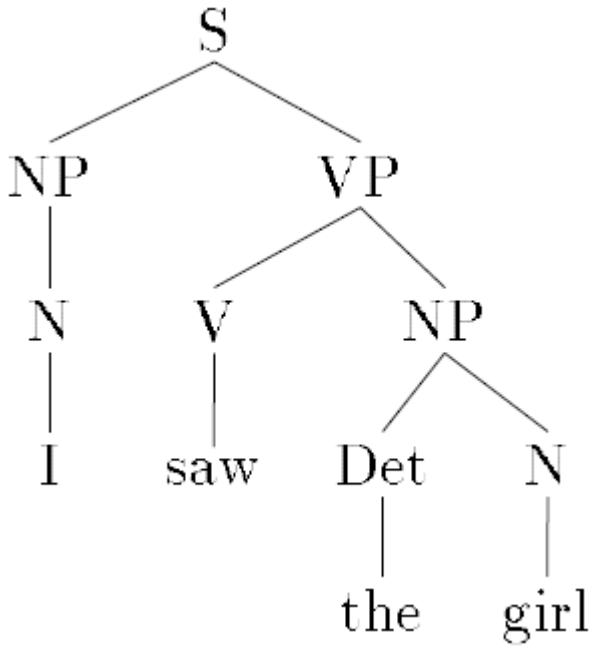
     Représentation de l'arbre syntaxique bien connu.
  2. **Structure fonctionnelle (f-structure)**

     Représentation des différents features associés aux constituants.

Grammaires génératives - LFG

C-structure

- Constitue l'arbre syntagmatique
- Exprimé par un ensemble de règles



Grammaires génératives - LFG

F-structure

- Constitue l'ensemble des relations sous-jacentes d'une grammaire impliquée
- Relations extraites sous forme d'annotations matricielles

$$\begin{bmatrix} \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'pro'} \\ \text{PERS} & 1 \\ \text{NUM} & \text{SG} \end{bmatrix} \\ \text{TENSE} & \text{PAST} \\ \text{PRED} & \text{'see}\langle(\uparrow \text{SUBJ}), (\uparrow \text{OBJ})\rangle\text{'} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'girl'} \\ \text{DEF} & + \\ \text{PERS} & 3 \\ \text{NUM} & \text{SG} \end{bmatrix} \end{bmatrix}$$

Grammaires génératives - LFG

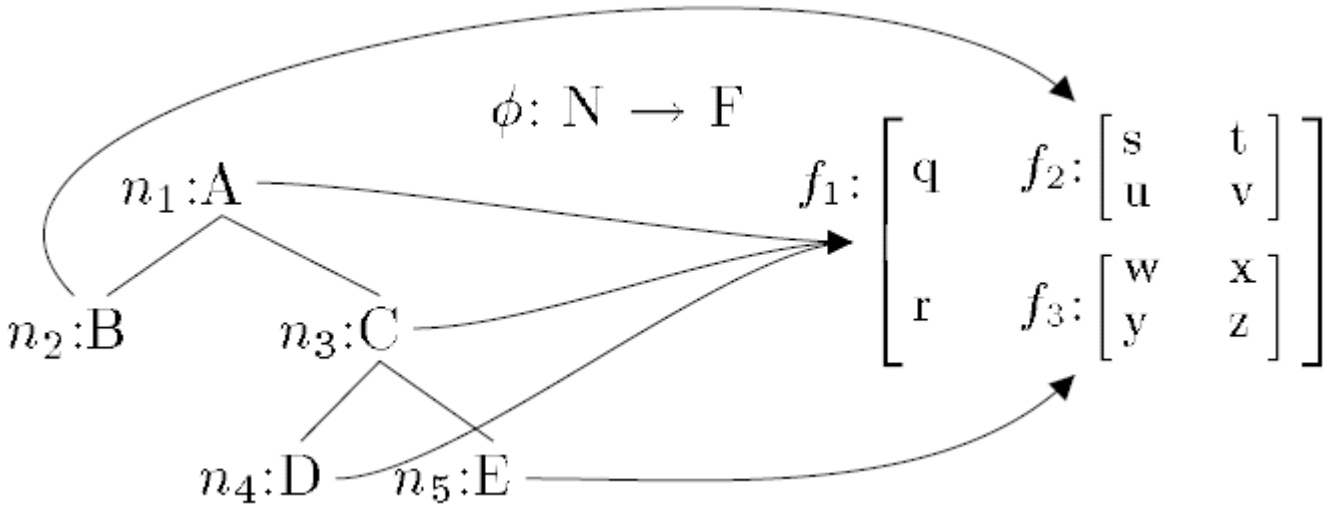
Features (Dalrymple, 2006)

	Feature	*Value*
Person	PERS	1, 2, 3
Gender	GEND	MASC, FEM, ...
Number	NUM	SG, DUAL, PL, ...
Case	CASE	NOM, ACC, ...
Surface form	FORM	Surface word form
Verb form	VFORM	PASTPART, PRESPART, ...
Complementizer form	COMPFORM	Surface form of complementizer: THAT, WHETHER, ...
Tense	TENSE	PRES, PAST, ...
Aspect	ASPECT	F-structure representing complex description of sentential aspect; sometimes abbreviated, e.g., PRES.IMPERFECT
Pronoun type	PRONTYPE	REL, WH, PERS, ...

Grammaires génératives - LFG

- Récursivement, des fonctions hiérarchiques définissent la valeur des relations.
- $(f_1\ q) = f_2$, $(f_2\ s) = t$, $(f_1\ r) = f_3$, etc.

  Où f_i est une structure (matrice)



Grammaires génératives - LFG

- Exemple:

« David sneezed »

- (*f* PRED) = 'SNEEZE'
- (*f* TENSE) = PAST
- (*f* SUBJ) = *g*
- (*g* PRED) = '*DAVID*'

$$f:\begin{bmatrix} \text{PRED} & \text{'SNEEZE}\langle\text{SUBJ}\rangle\text{'} \\ \text{TENSE} & \text{PAST} \\ \text{SUBJ} & g:\begin{bmatrix} \text{PRED} & \text{'DAVID'} \end{bmatrix} \end{bmatrix}$$

où PRED = forme sémantique, TENSE = temps du verbe, SUBJ = Sujet

Grammaires génératives - LFG

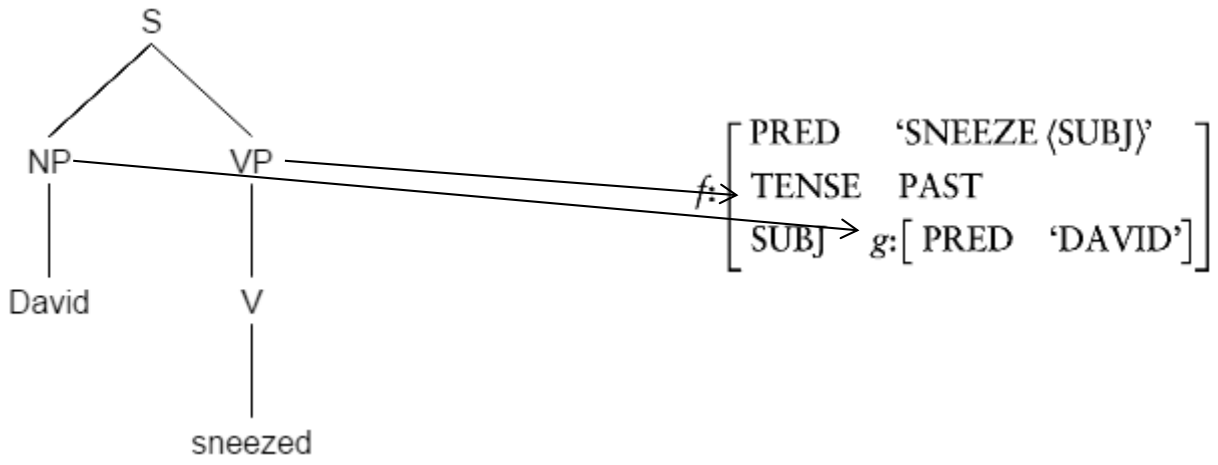
- Φ: Fonction de relation propre à chaque nœud de l'arbre permettant le couplement d'un *POS* à une unité de la f-structure. Exemple formel:

$$
\begin{array}{ll}
(\phi(\mathrm{M}(n_2))\ \mathrm{q}) = \phi(n_2) & \mathrm{M}(n_2) = n_1 \\
(\phi(n_2)\ \mathrm{s}) = \mathrm{t} & \\
(\phi(n_5)\ \mathrm{y}) = \mathrm{z} & \\
\phi(\mathrm{M}(n_3)) = \phi(n_3) & \mathrm{M}(n_3) = n_1 \\
\phi(\mathrm{M}(n_4)) = \phi(n_4) & \mathrm{M}(n_4) = n_3 \\
(\phi(\mathrm{M}(n_5))\ \mathrm{r}) = \phi(n_5) & \mathrm{M}(n_5) = n_3 \\
\mathrm{etc.} &
\end{array}
$$

où n_i = *POS* i et M(n_i) = Parent de n_i

Grammaires génératives - LFG

- Φ: Exemple «moins formel»:



À noter que la fonction Φ peut être une relation multiple → unique

Grammaires génératives - HPSG

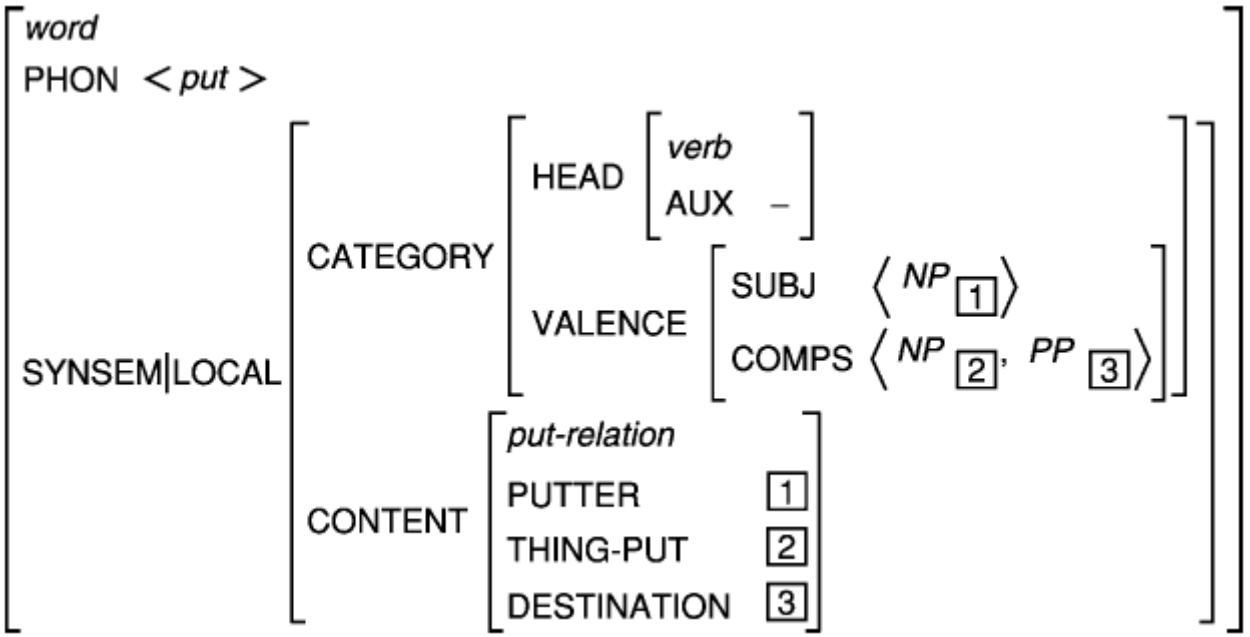
- Deux composants essentiels définissent les HPSG:
  1. Structure de représentation explicite des catégories de la grammaire (non-terminaux)
  2. Ensemble de contraintes décrivant la généralisation linguistique des catégories sous forme de série matricielle.
     - Ces matrices possèdent les propriétés d'une AVM (Attribute Value Matrix) et sont regroupées dans un ensemble nommé SYNSEM (Syntaxic-Semantic).

Grammaires génératives - HPSG

- Composition du SYNSEM
  - Entrée lexical: le terme d'une phrase
  - PHON: une liste de chaînes de mot reproduisant la phonétique du terme
  - CATEGORY: information morphosyntaxique du terme
  - CONTENT: information sémantique du terme
  - VALENCE: ses liaisons lexicales potentielles et ses rôles sémantiques
    - SUBJ, COMPS
  - HEAD: relie les structures d'une même lignée syntaxique
  - TYPE: word/phrase (e.g. « Robert » ou NP

Grammaires génératives - HPSG

- Exemple (Levine et Meurers, 2006)



Grammaires génératives - HPSG

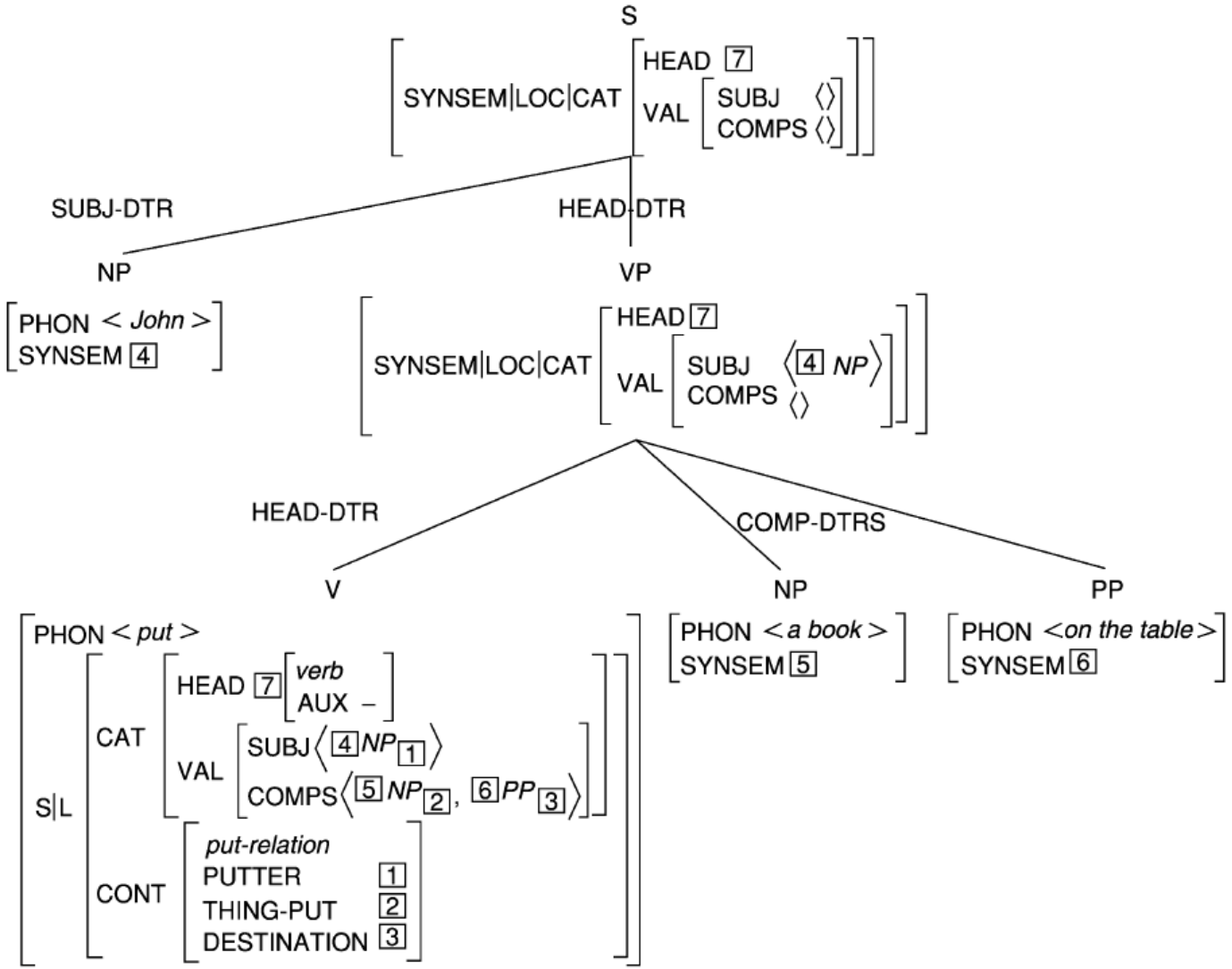
- Principes primaires des HPSG
  - Head Feature Principle

  S'assure que les jetons d'identification des structures possédant une propriété HEAD sont identiques.

  - Valence Feature Principle

  S'assure que les valeurs des propriétés VALENCE regroupe l'ensemble des compositions incomplètes.

Grammaires génératives - HPSG

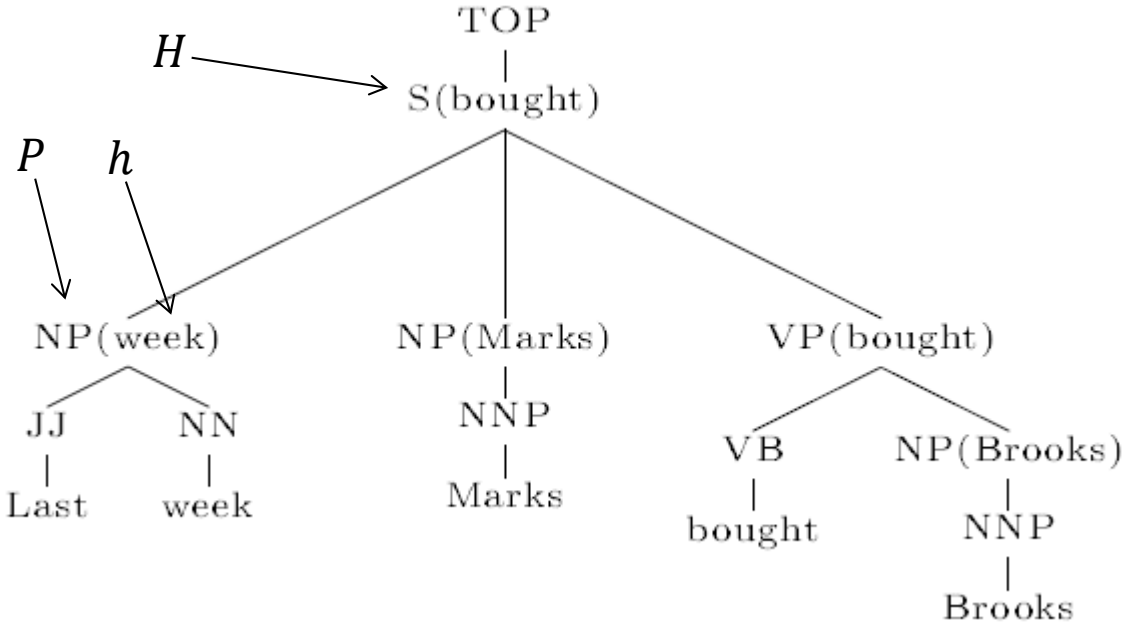


Grammaires génératives - Autres

- (Collins, 1997)
  - Les PCFG de (Magerman, 1995) et (Jelinek et al., 1994) estiment directement P($T|S$).
  - Présentation de trois modèles génératifs inspirés des PCFG lexicalisés:
    1. Version générative de base
    2. Distinction des compléments
    3. Gap Feature (The car that she bought?)

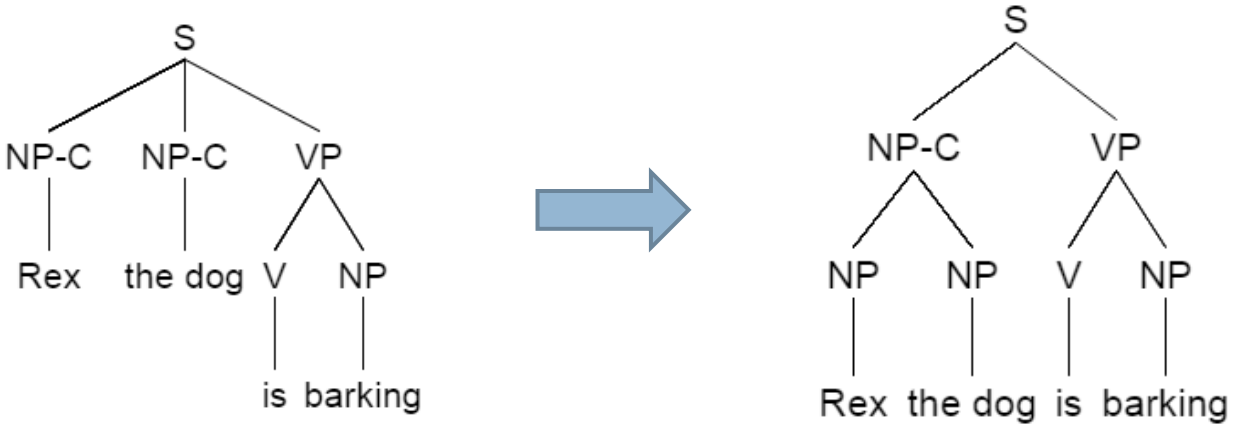
Grammaires génératives - Autres

- Modèle 1 (hypothèses markoviennes d'ordre 0)
  - On génère le constituent de tête de phrase: $\mathcal{P}_H(H|P,h)$
  - On génère les modificateurs de droite: $\prod_{i=1..m+1} \mathcal{P}_R(R_i(r_i)|P,h,H)$
  - On génère les modificateurs de gauche: $\prod_{i=1..m+1} \mathcal{P}_L(R_i(l_i)|P,h,H)$
    - $m+1$ = non-terminal d'arrêt (STOP)



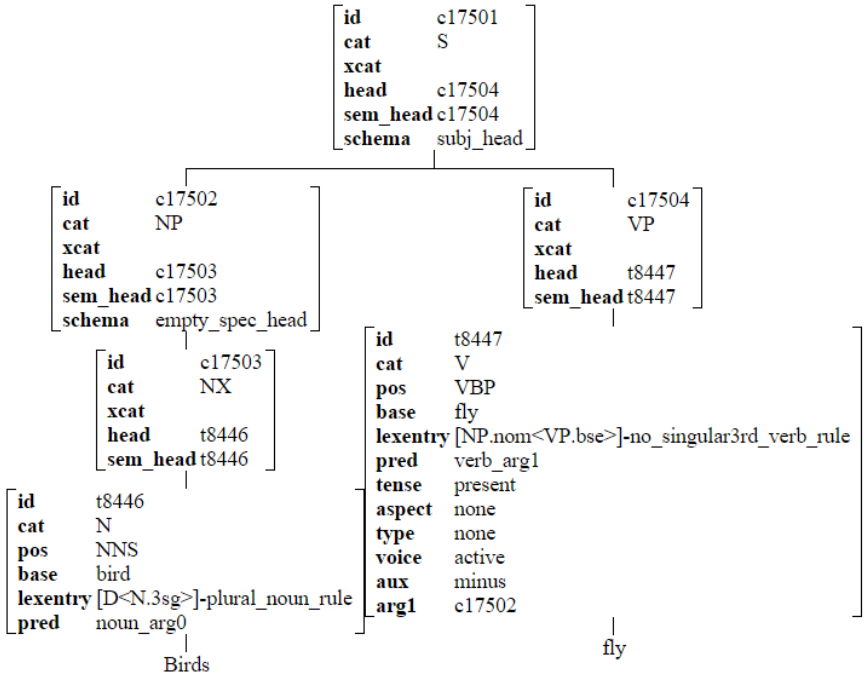
Grammaires génératives - Autres

- Modèle 2
  - Les compléments sont identifiés au sein de la grammaire avec la notation -C.
  - On ajoute les probabilités qu'un constituent de droite ou de gauche soit un complément: $\mathcal{P}_{rc}(RC|P,H,h)$ et $\mathcal{P}_{lc}(LC|P,H,h)$
  - La probabilité $\mathcal{P}_{rc} \wedge \mathcal{P}_{lc}$ est réduite dans le cas de deux NP adjacents.
    - Ceci a pour effet d'unifier les séquences de phrase composant un même complément (sujet ou objet).



Grammaires génératives - Parseurs

- HPSG
  - Enju
    - Université de Tokyo
    - http://www.nactem.ac.uk/enju/
    - Précision de 90% pour les relations prédicat-argument



Grammaires génératives - Parseurs

- LFG
  - National Centre for Language Technology Probabilistic Parser
  - Dublin City University

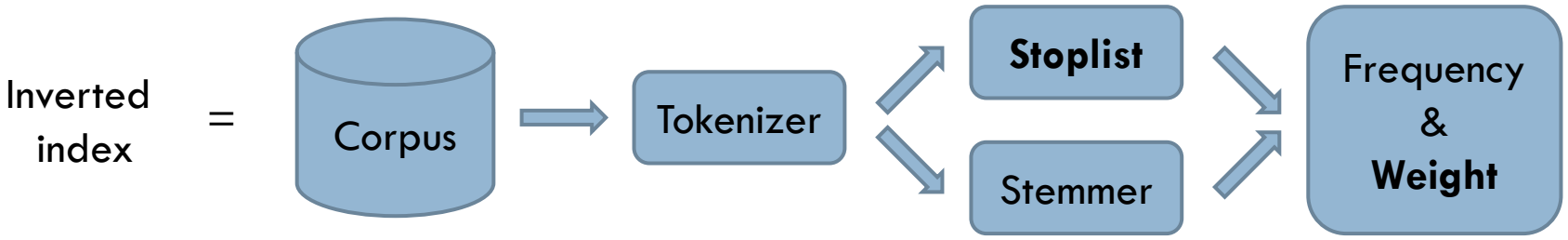
```
subj : pred : mary
       num : sg
       pers : 3
pred : like
tense : pres
num : sg
pers : 3
obj : spec : det : pred : the
      pred : park
      num : sg
      pers : 3
      relmod : topicrel : pform : in
                          obj : pred : pro
                                pron_form : which
               subj : pred : john
                      num : sg
                      pers : 3
               pred : walk
               tense : pres
               num : sg
               pers : 3
               adjunct : pform : in
                         obj : pred : pro
                               pron_form : which
               resolved : topicrel
```

Processus de recherche

Indexation et pondération

Indexation

- Stockage des termes d'un ensemble de documents dans un index inversé.
- Les phases stopper et weighting sont relativement lié aux grammaires…



Indexation

Stoplist

- Liste de termes à ignorer durant l'indexation.
- Généralement composée de mots sans portée sémantique significative (e.g. a, the, and, at, etc.).

Weighting

- Mesure de l'importance d'un terme de requête dans une collection de documents.
- Le modèle de pondération le plus connu pour l'indexation est TF-IDF.

Pondération

- Regroupe les critères de pertinence pour le classement des documents selon la requête d'un usager.
- Plusieurs modèles de pondération peuvent être employés:
  - Modèle de langue statistique (n-gramme)
  - Modèle booléen
  - Modèle probabiliste (TF-IDF, BM25):

$$tf_{ik}\ idf_k = freq(t_k, D_i) \times log(|D|, |D|_k)$$

Indexation et pondération

Participation des grammaires

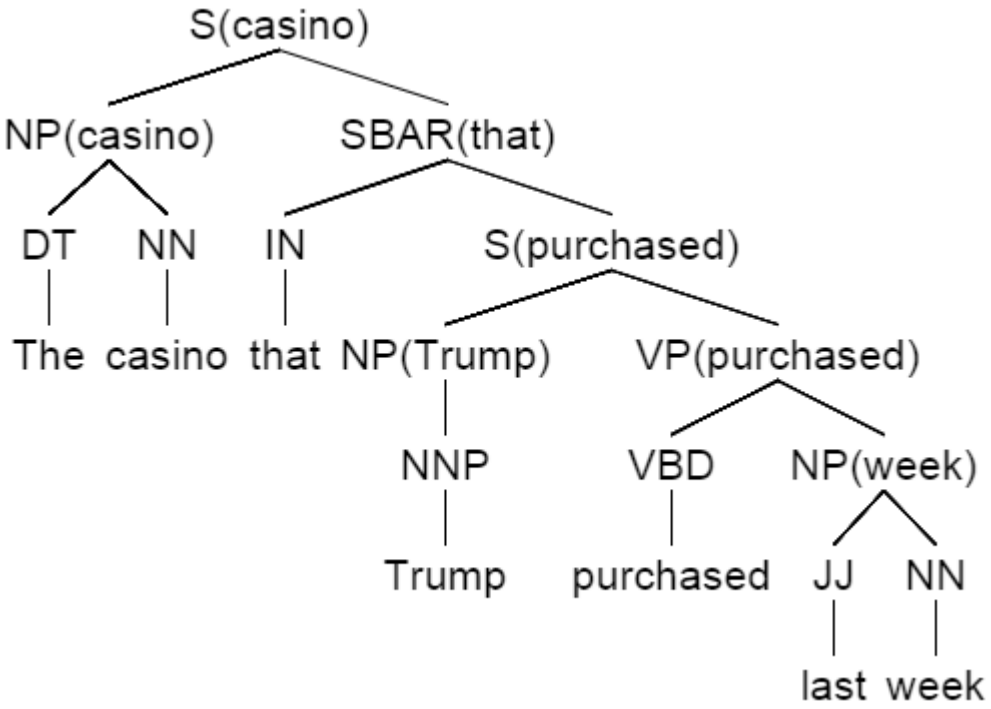
- Utilisation d'un paramètre (*p*) de catégorisation des termes déterminant l'importance des parties d'une requête suivant les propriétés des grammaires génératives de (Collins, 1997).
  - Un terme de requête rattaché à une tête lexicale en deçà du niveau *p* dans *T* obtient un score inférieur durant l'indexation.
  - Un terme de requête n'étant représenté par aucune tête lexicale a plus de probabilité d'être listé parmi les stop words…

Indexation et pondération

Exemple $(p = 2)$:

«The casino that Donald Trump purchased last week»

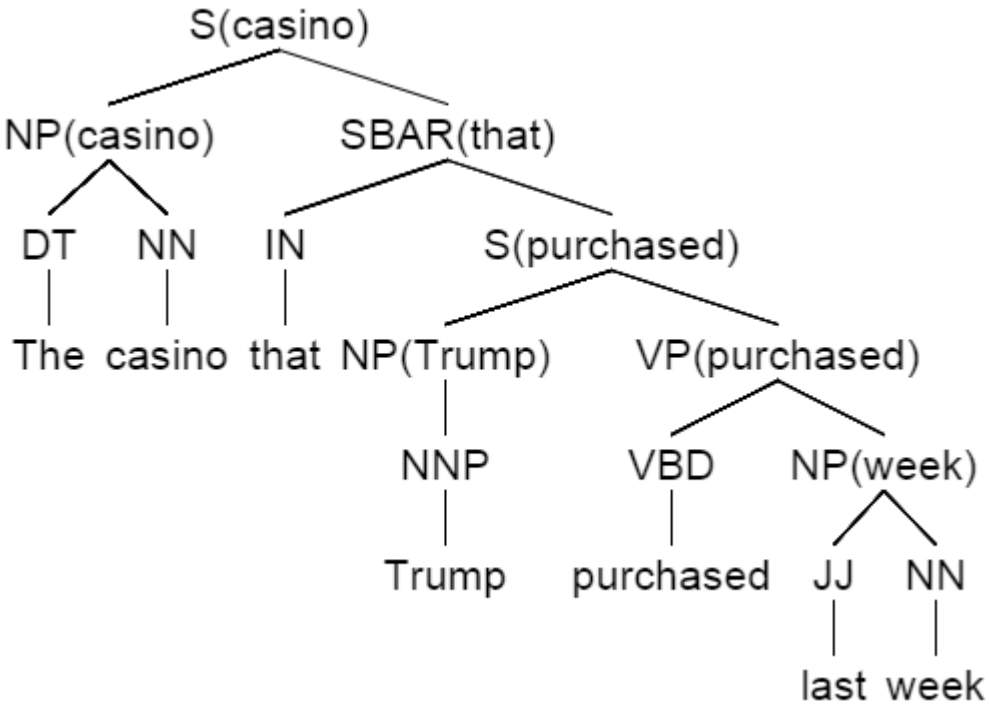
- casino > purchased > Trump > week
- «that» est filtré par le stopper



Indexation et pondération

Exemple $(p = 2)$:

- $Si\ hauteur(t_i) <\ p\ alors\ z(t_i) = \lambda\ sinon\ z(t_i) = (1 -\ \lambda)$
- $tf_{ik}\ \ idf_k = z(t_i) \times freq(t_k, D_i) \times log(|D|, |D|_k)$
  - $z(casino) = \lambda$
  - $z(purchased) = \lambda$
  - $z(Trump) = \lambda$
  - $z(week) = 1 - \lambda$



Processus de recherche

Expansion de requêtes

Expansion de requêtes

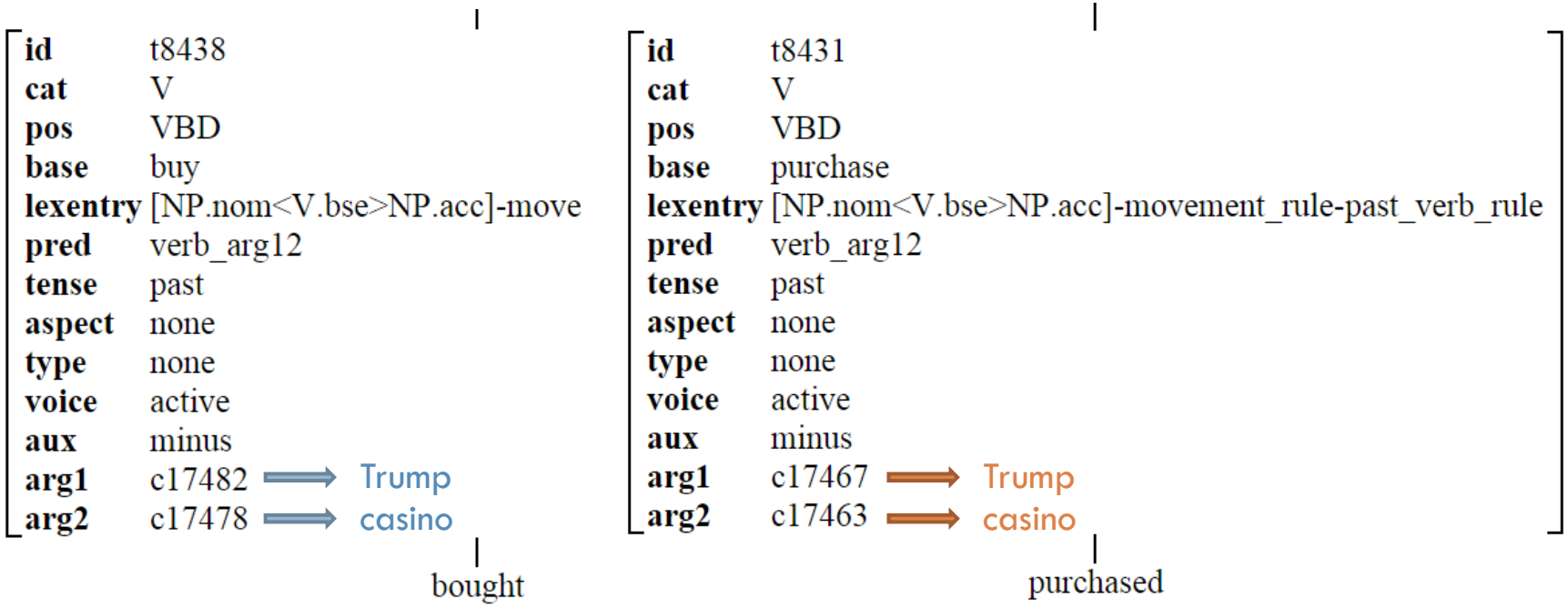
- Augmentation d'une requête par des termes jugés équivalents.
- Beaucoup d'approches répertoriées:
  - Utilisation d'un thésaurus (synonymes, hyponymes, hyperonymes, méronymes, etc.)
  - Cooccurrences de termes dans un même document
  - Pseudo-Relevance Feedback
    - *k* meilleurs documents possède une plus grande probabilité de contenir des termes supplémentaires pertinents dans une fenêtre *w* entourant un terme q_i.

Expansion de requêtes

Participation des grammaires

- Deux verbes partageant les mêmes compléments sont des candidats à l'expansion.

Enju Parser: «The casino that Donald Trump bought/purchased last week»



Contribution

Travaux

Travaux (Voorhees, 1994)

Query Expansion using Lexical-Semantic Relations

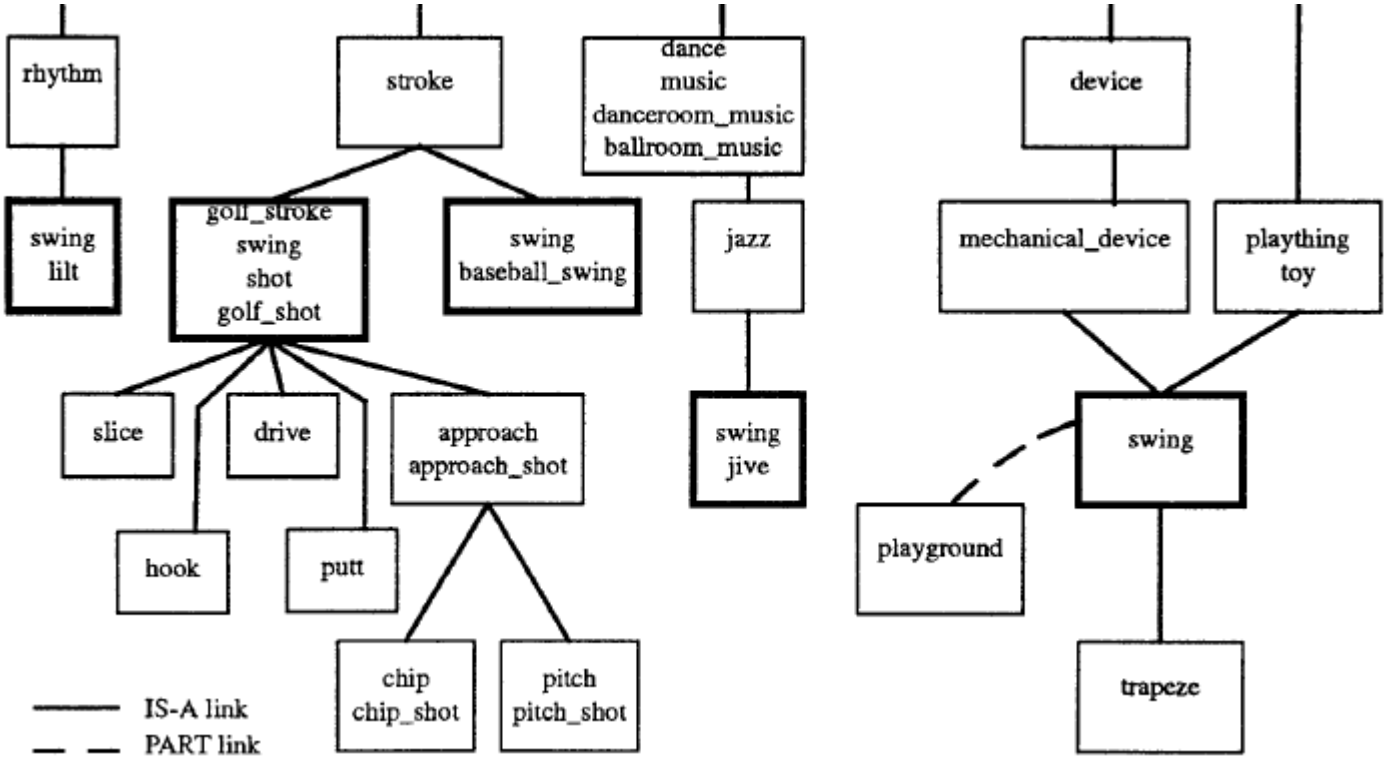
- Étude de l'utilité de l'expansion de requêtes par l'usage de relations lexico-sémantiques.
- TREC
- WordNet (source des relations lexico-sémantiques)
- Expérimentation manuelle et tentative d'automatisation du processus de sélection de termes.
- Difficultés
  - Un concept est souvent décrit par plus d'une expression.
  - Une expression décrit parfois plusieurs concepts.
  - Comment sélectionner les termes automatiquement?

Travaux (Voorhees, 1994)

- TREC
  - 50 requêtes (topics)
    - Descriptions
    - Concepts
    - Domaines
  - 742 000 documents (journaux, documents techniques, etc.)

Travaux (Voorhees, 1994)

- WordNet
  - Relations conceptuelles
    - Synonymes, hyponymes, hyperonymes, méronymes



Travaux (Voorhees, 1994)

- Expérimentation
  1. Manuelle
     - Cancer → {cancer} {skin_cancer} {pharmaceutical}
     - Pour chaque requête
       - Création de vecteurs de termes associés à un *ctype*
         - Synonymes
         - Synonymes + Descendants *is-a*
         - Synonymes + Descendants *is-a* + Parents *is-a*
         - Synonymes + Concepts directement liés
  2. Automatique
     - Un terme qui apparaît dans plus de N documents n'est pas candidat à l'expansion.
     - Un terme qui n'est pas relié à au moins deux termes de la requête originale n'est pas candidat à l'expansion.

Travaux (Voorhees, 1994)

- Résultats

	Synonymes	Synonymes + Desc. *is-a*	Synonymes + Desc. *is-a* + Parents *is-a*	Synonymes + rel. directes
Manuelle	+1.5%	+1.5%	+1.7%	+1.2%

	N=70 000 Length=1	N=70 000 Length=2	N=35 000 Length=1	N=35 000 Length=2
Automatique	-0.5%	-0.1%	+0.3%	+0.7%

Travaux (Oliveira et al., 2007)

Query Translation for Cross-Language Information Retrieval by Parsing Constraint Synchronous Grammar

- Différentes approches pour la traduction en recherche d'information
  - Usage d'un thésaurus
    - Les vocabulaires sont trop imposants et les mots contiennent parfois plusieurs traductions
  - Machine Translation à base de règles
    - Grand nombre de règles à maintenir
  - Machine Translation statistique
    - Dépend entièrement du corpus parallèle (bitexte)

Travaux (Oliveira et al., 2007)

- Mise en application de CSG (Constraint Synchronous Grammar) pour la traduction de requêtes courtes.
  - Variante d'un SG (Synchronous Grammar)
    - Exprime des traits lexicales et syntaxiques à la manière d'un HPSG
    - La qualité d'une traduction passe par les contraintes de la grammaire.
  - Chinois → Portuguais

Travaux (Oliveira et al., 2007)

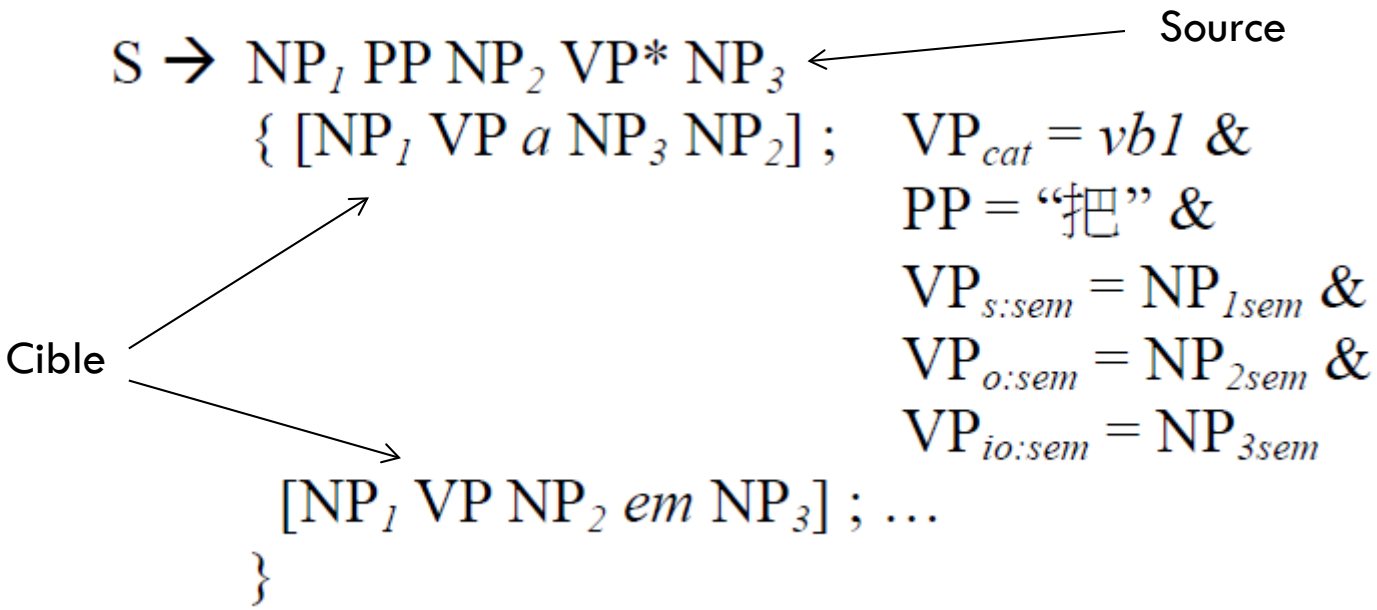
Mise en œuvre

- Segmentation des mots de la requête source (chinois)
- Tagging
- Analyse guidée par les CSG et un parseur LR pour déterminer le patron séquentiel des mots cibles (portuguais)
- Traitement de la requête par un engin de recherche monolingue.

Travaux (Oliveira et al., 2007)

CSG

- Ensemble de règles de production qui décrivent le patron de la séquence de parts of speech que doit adopter la requête cible.
- La règle qui satisfait toutes les conditions est choisie par le traducteur.

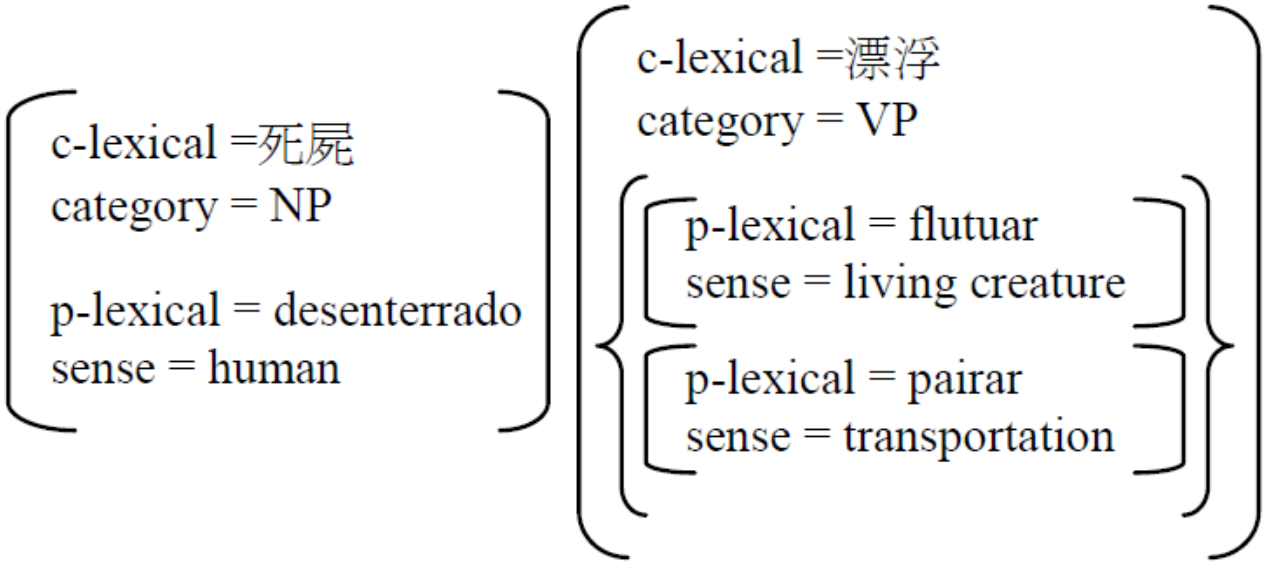


* = head

Travaux (Oliveira et al., 2007)

CSG – Feature Descriptors (FD)

- Encodés sous la forme de matrices AVM
- Chaque feature est associé à un poids initial.
- Si l'unification est possible entre deux FD, le poids est augmenté de 1.
- Si l'unification échoue mais que le sens d'un terme est l'hyponyme de l'autre, le poids est augmenté de 0.5.



Travaux (Koster et al., 1996)

AGFL Grammars for full-text Information Retrieval

- Élaboration de grammaires AGFL pour contrer certains problèmes reliés à la recherche d'information:
  - Construction syntaxique déficiente
  - Fautes d'orthographe
  - Constructions idiomatiques
  - …
- Quelques propriétés recherchées pour une grammaire efficace:
  - Bonne précision, couverture large, faible ambiguïté

Travaux (Koster et al., 1996)

AGFL

- Grammaire hors-contexte avec features
- Description morphosyntaxique du langage naturel
- Composée de règles ou meta-règles (*affix*)
  - Non-terminaux / Terminaux
    - numb :: SING; PLUR
    - person :: FIRST; SECOND; THIRD
  - Domaine
    - Groupe de productions de ses terminaux

Travaux (Koster et al., 1996)

Exemple:

```
numb :: SING; PLUR.
person :: FIRST; SECOND; THIRD.
to be (SING, FIRST) : "am".
to be (SING, THIRD) : "is".
to be (PLUR, FIRST|THIRD): "are".
to be (numb, SECOND) : "are".
simple sentence :
  pers pron (numb, pers), to be (numb, pers), adjective.

pers pron (SING, FIRST) : "I".
pers pron (PLUR, FIRST) : "we".
pers pron (numb, SECOND) : "you".
pers pron (SING, THIRD) : "he"; "she"; "it".
pers pron (PLUR, THIRD) : "they".

adjective : "great".
```

Travaux (Koster et al., 1996)

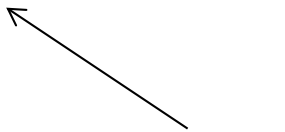
Particularités de la grammaire

- Principes de stratification
  - Ordonnancement des parties d'une phrase analysée sous forme de classes

```
utterance:
    sentence, [sentence]!
    noun phrase [sentence]!
    wild card, [sentence].
```

  - Mécanisme de pénalité
    - Formalité de pondération (pénalité + 1 → poids - 1)

```
UNKNOWN PRESPART: $PENALTY, $MATCH(".*ing").
```



Conclusion

HPSG

LFG

PCFG

AGFL

CSG

- Possibilités
  - Désambiguïsation
    - Diminution du bruit
  - Apport sémantique
    - Rappel
  - Apport syntaxique
    - Précision

Merci