Exploitation de ressources linguistiques en génération de texte multilingue

Florie Lambrey

Observatoire de Linguistique Sens-Texte

Plan

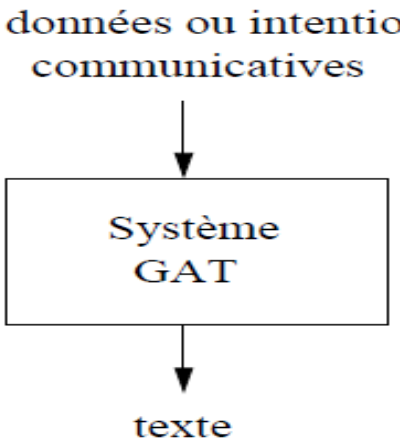
- Introduction
- Architecture d'un système de GAT
- Ressources linguistiques en GAT(M)
- Système MARQUIS
- Conclusion

Plan

- **Introduction**
- Architecture d'un système de GAT
- Ressources linguistiques en GAT(M)
- Système MARQUIS
- Conclusion

Introduction

- Sous domaine du NLP, mélange entre IA et LC
- Essor depuis les 70s et 80s
- Enjeux :
  - Théoriques : interaction homme-machine, représentation et conversion de contenus
  - Pratiques : présentation de l'information, automatisation ou assistance des tâches de rédaction
- Données :
  - Input : contenu non textuel
  - Output : équivalent linguistique



Introduction

- NLG se distingue du NLU (Reiter et Dale, 2000):
  - '**Hypothesis managment**' → déterminer la bonne interprétation étant donné un contexte donné
  - '**Choice**' → partir d'une représentation et choisir la bonne formulation

- Faux générateurs de texte : technique du Mail Merge (pré-enregistrements) :
  - Messages d'erreur
  - Communications STM
  - Réponses automatiques ...

→ savoir identifier les besoins des utilisateurs

Plan

- Introduction
- **Architecture d'un système de GAT**
- Ressources linguistiques en GAT(M)
- Système MARQUIS
- Conclusion

Architecture d'un système de GAT

quoi dire : forme logique du message à transmettre

comment le dire : formulation linguistique à partir de la forme logique

Module	*Content task*	*Structure task*
Document planning	Content determination	Document structuring
Microplanning	Lexicalisation; Referring expression Generation	Aggregation
Realisation	Linguistic realisation	Structure realisation

Figure 3.1 Modules and tasks.

Plan

- Introduction
- Architecture d'un système de GAT
- **Ressources linguistiques en GAT(M)**
- Système MARQUIS
- Conclusion

Ressources linguistiques en GAT(M)

Approche symbolique

création de ressources encodant les informations linguistiques sensibles pour la GAT :

- Lexiques : inventaire des unités d ela langue
- Grammaires : agencement de ces unités

→ formalisation de ces ressources :

- Lexical Functional Grammars
- Grammaires d'arbres adjoints
- ...

Ressources linguistiques en GAT(M)

Problèmes sur l'utilisation de ces ressources (et la conception des systèmes de GAT) :

- Spécialisation lexicale
- Monolingue
- interopérabilité difficile
- Chronophage et coûteux
- Expertise

→ Comment exploiter ces ressources en vue de créer un système de GATM?

Plan

- Introduction
- Architecture d'un système de GAT
- Ressources linguistiques en GAT(M)
- **Système MARQUIS**
- Conclusion

Système MARQUIS

Multimodal **A**i**R** **Qu**ality **I**nformation **S**ervice

Langues traitées : anglais, catalan, portugais, finnois, français, allemand et polonais

Cadre théorique : Théorie Sens-Texte

Deux avantages :

- Architecture modulaire qui représente le fonctionnement de la langue (analyse et production)
- Traitement des collocations (et des paraphrases)

Système MARQUIS

Architecture d'un modèle TST

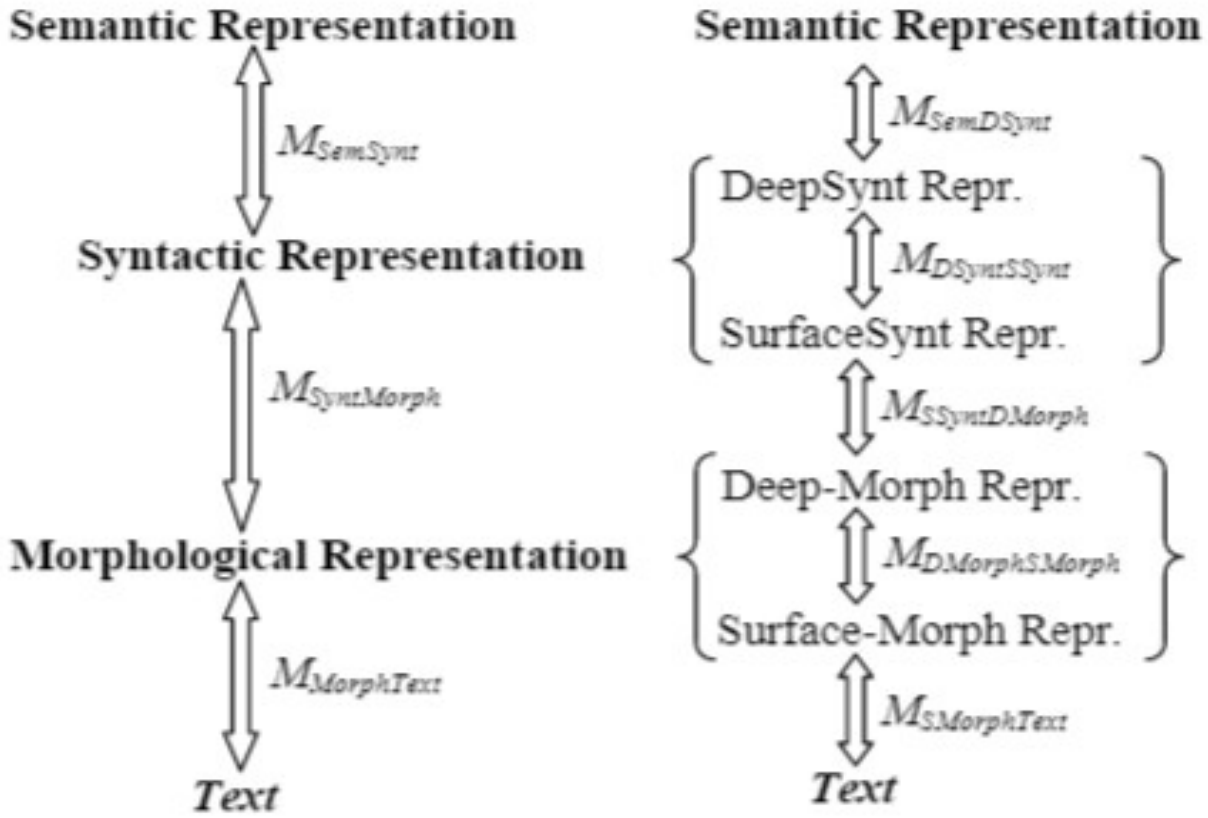


Figure 1: Meaning-Text Linguistic Model

Système MARQUIS

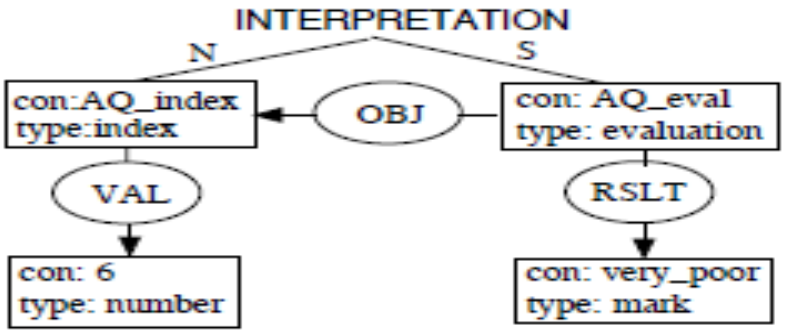
Plateforme MATE (Meaning Text Environment)

- Représentation des niveaux en différents types de graphes
- Création de règles de correspondance entre ces niveaux
- Les règles vont faire appel aux informations contenues dans les lexiques

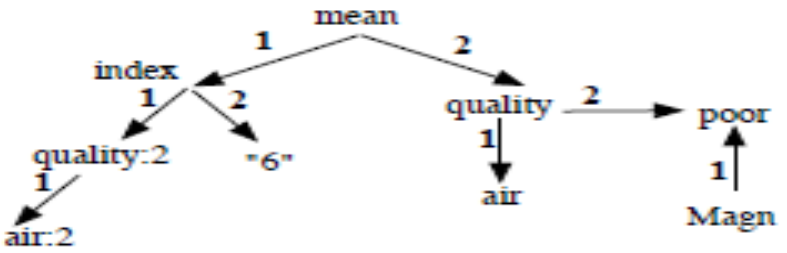
air quality index / air quality	Values of the air quality index (e.g., 'good', 'satisfactory', 'bad', etc.); the number and exact labels of the air quality index values depend on the index used. In MARQUIS, five different country-specific indexes were used.
individual air pollutants	'particulate matter' ('dust particles', 'PM10' / 'PM2.5'), ozone, nitrogen dioxide, sulphur dioxide, etc.
units	µg/m³, hour
concentration	e.g. 175 µg/m³
thresholds	'information threshold' , 'alarm threshold', etc.
forecasts	'forecast index', 'forecast index', etc.
justification / forecast motivation	weather, traffic, etc.

Système MARQUIS

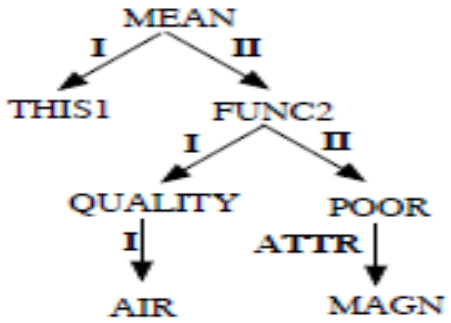
1. Conceptual Structure



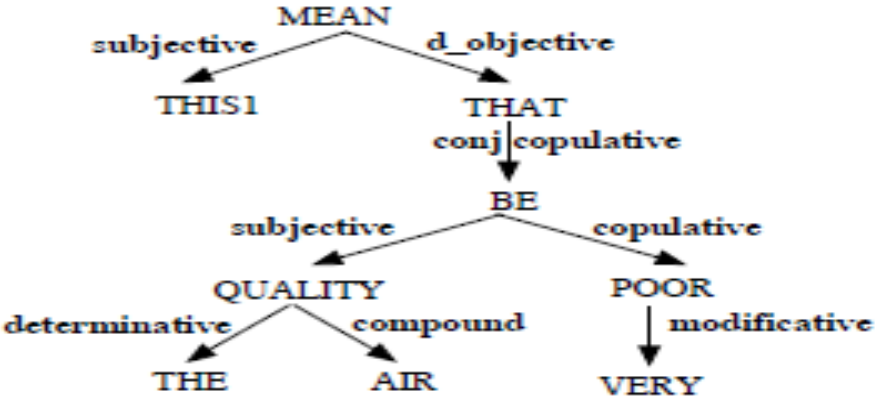
2. Semantic Structure



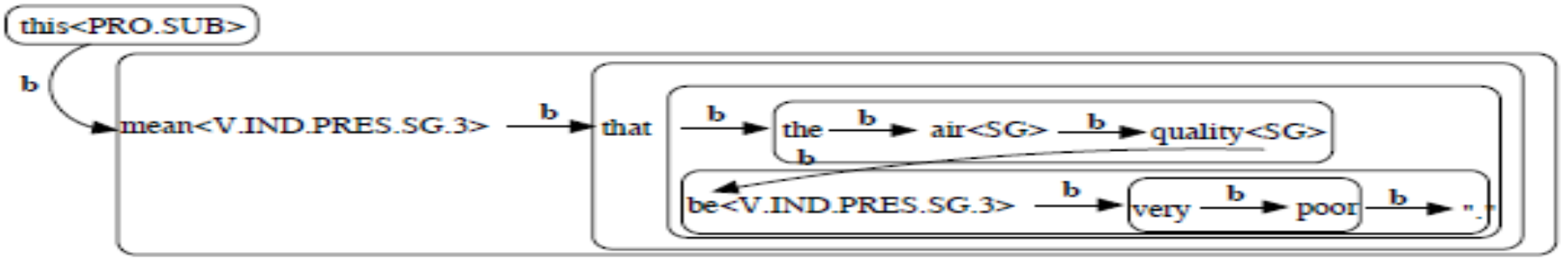
3. Deep–Syntactic Structure



4. Surface–Syntactic Structure



5. Deep–Morphological Structure



Système MARQUIS

Traitrement de collocations en Sens-Texte

Hypothèses linguistiques:

- Il existe des liens de **dépendance** sémantique et lexicale entre les unités de la langue
- Cela s'exprime par des choix combinatoires restreints, ce qu'on appelle des **collocations (étape de lexicalisation)**
- Ces liens représentent des concepts récurrents dans plusieurs langues et peuvent se **factoriser (fonctions lexicales)**

  ex. Peur *bleue*; remercier *chaleureusement*; brouillard *dense*; *grièvement* blessé vs. *Gravement* malade

- (lien = **intensification**)

$$\text{Sém(FL)} = \frac{L1}{L2} = \frac{L3}{L4}$$

Système MARQUIS

Fonctions lexicales : outil de modélisation des collocations

Table 1: Patrons de fonctions lexicales

Patron	**Fonctions lexicales**	**Exemples**
ATTR	Magn	peur *bleue*
	Ver	appareil *exact*
Verbes Support	Oper	*avoir* froid, *rendre* fou
	Func	la neige *tombe*
	Labor	*entourer* quelqu'un de soins
Noms Supports	Figur	*rideau* de fumée
Prep	Locab	*en plein* front
	Locin	*au sein du* personnel
Adj/A	A2	*couvert de* mépris
	Adv1	*avec* joie

Système MARQUIS

Exemple de règle SemR <=> DsyntR (lexicalisation de Func)

Contexte de gauche
(structure en entrée)

concentration{
  1-> **100 ug/m** {}
}
S*{
  sem=S
  concentration
  100 ug/m
}

<=>

Contexte de droite
(structure cible)

Etre { <=> concentration
  split=top
  dlex=Func+1
  dpos=V
  I-> **concentr**. { <=> concentr.
  split=bottom
  dlex =(concentration).(lex)
  dpos = lexicon::(concentr).(dpos) }
  II->**100 ug/m**{ <=> 100 ug/m
  dpos= (lf::((Func+1)).(gp).(1)).(dpos) }
}

Conditions d'application de la règle
lexicon::(concentration).(Func+1)
Etre.dpos=V or (not Etre.dpos)

Système Marquis

```
concentration : noun {
  pos = N
  countable = yes
  gp = {
     1 = I
     2 = II
     I = {  dpos = N    rel = compound det = no }
     I = {  dpos = N    rel = noun_completive prep = of  det = no }
     II = { dpos = Num  prep = of }
     II = { dpos = Adj  rel = modificative }
     II = { dpos = Adv  rel = modificative }
  }
  Magn = high
  AntiMagn = low
  Adv1 = in             // "we found ozone in a concentration of 180 µg/m3"
  Func2 = be1           // "the concentration (of ozone) is 180 µg/m3"
  IncepFunc2 = reach    // "the concentration (of ozone) reached 180 µg/m3"
  IncepOper1 = reach    // "ozone will reach a concentration of 180 µg/m3"
}
```

Système MARQUIS

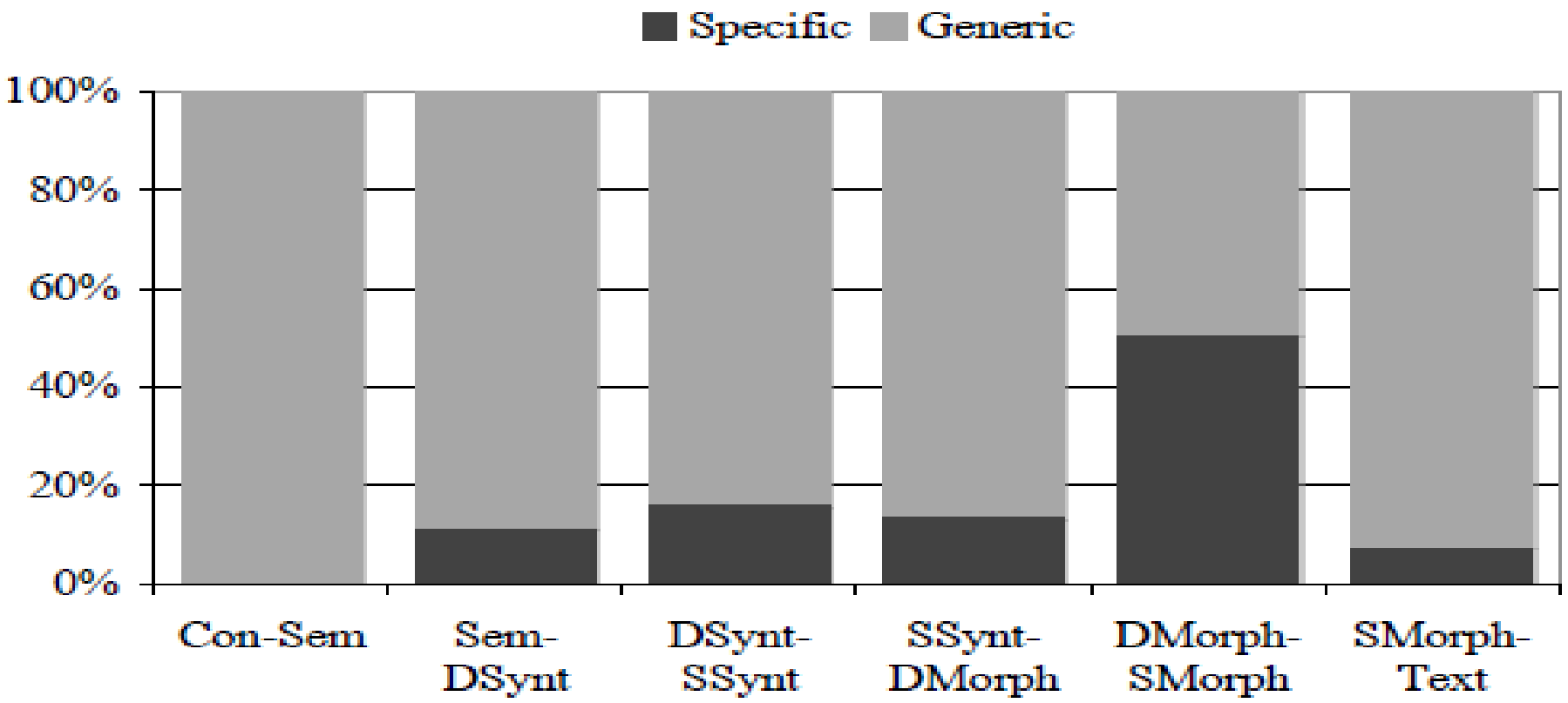


Figure 2: Average generic / language-specific rule ratio by module

Conclusion

- L'utilisation de ressources linguistiques permet de générer des phénomènes linguistiques précis
- L'approche Sens-Texte permet de capter des régularités de la langue et de générer des paraphrases
- D'autres approches existent pour la NLG mais ne sont pas mentionnées ici
- Des projets existent (au moins un!) dont le but est de rendre les systèmes de GATM plus génériques (GÉCO)

Merci de votre attention

Bibliographie

- Lareau, F., Dras, M., Börschinger, B., & Dale, R. (2011, December). Collocations in multilingual natural language generation: lexical functions meet lexical functional grammar. In Proceedings of the Ninth Australasian Language Technology Workshop (ALTA'11) (pp. 95-104).
- Lareau, F., & Wanner, L. (2007). Towards a generic multilingual dependency grammar for text generation. In T.H. KING & E.M. BENDER, Eds., Proceedings of the GEAF07 Workshop, p. 203-223, Standford:CSLI.
- Reiter, E., & Dale, R. (2000). Building Natural Language Generation Systems. Cambridge University Press.
- STEINLIN, J. (2003). Générer des collocations, mémoire de master. Paris: Université de Paris VII-Denis Diderot.