

# Génération automatique de texte (NLG)

## État de la recherche et des applications



Serge Bibauw

Chercheur en linguistique appliquée  
Centre de traitement automatique du langage  
**Université catholique de Louvain**

Audaxis Autumn Talk  
Bruxelles, 10 octobre 2017



# Quel est le point commun entre...

millions  
évènements  
mineurs



intérêt local  
et immédiat



matches de football

embouteillages

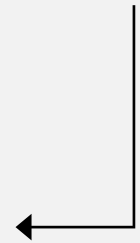
tremblements de terre

perturbations dans les  
transports en commun

mouvements boursiers



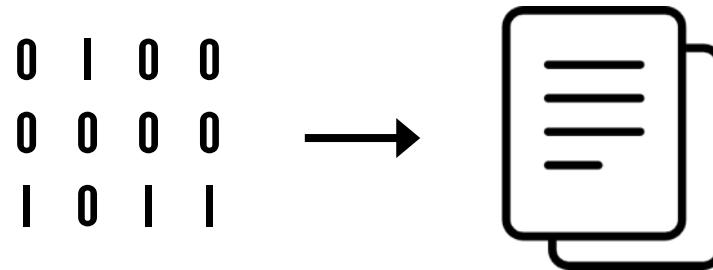
0	1	0	0
0	0	0	0
1	0	1	1



Accident sur le  
ring du Bruxelles :  
Tremblement de terre de  
magnitude 3  
à Spa

Le RFC Tournai  
victorieux face  
au Stade Brainois

# Génération automatique de texte (NLG)



**data-to-text**

# Génération automatique de texte (NLG)

**text-to-text**



**data-to-text**

0	1	0	0
0	0	0	0
1	0	1	1



**image-to-text**



# Génération automatique de texte (NLG)

**text-to-text**



**data-to-text**

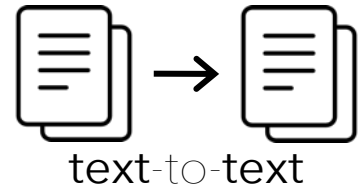
0	1	0	0
0	0	0	0
1	0	1	1



**image-to-text**



# Génération automatique à partir d'autres documents



## Fusion et **synthèse** de documents

Clarke & Lapata, 2010

## **Simplification** automatique

Siddharthan, 2014; Macdonald & Siddharthan, 2016

## Génération de **questions**

Brown et al., 2005; Rus et al., 2010

## Génération de **réponses** dans un système de dialogue (chatbot)

Rieser & Lemon, 2009; Serban et al., 2015

# Génération automatique de texte (NLG)

text-to-text



data-to-text

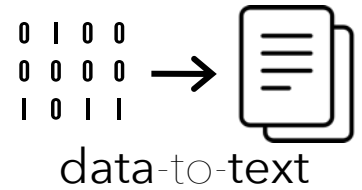
0	1	0	0
0	0	0	0
1	0	1	1



image-to-text



# Génération automatique de rapports et dépêches



## Dépêches **météo**

**FoG** (Goldberg et al., 1994)

Weather reports for offshore platform (Reiter et al., 2005)

## Mouvements **financiers**

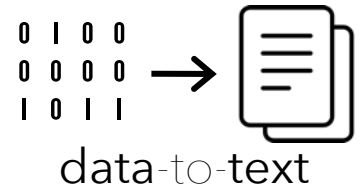
*Narrative Science pour Forbes*

## Descriptions d'**entreprises**

**FactSet Yseop**

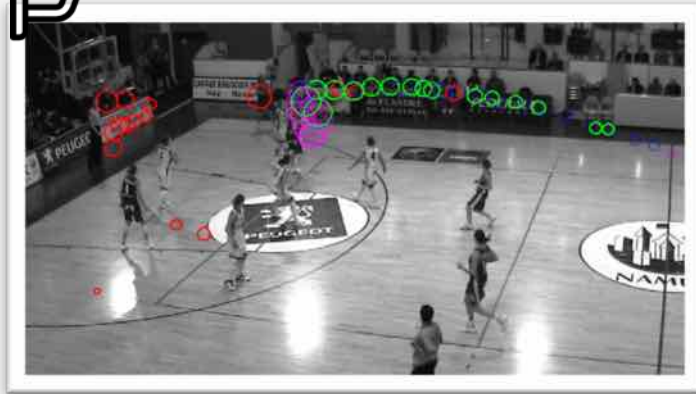


# Génération automatique de dépêches **sportives**

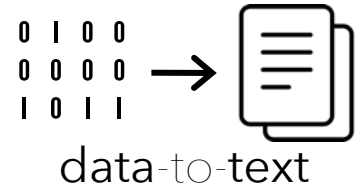


**IBM Slamtracker** : Wimbledon

**SPORTIC** (UCL/CENTAL, Acapela, ACIC SA) :  
compte-rendu sportif automatique (basket)



# Génération automatique à partir de données de senseurs



Suivi de **migrations d'oiseaux**

Siddharthan et al., 2013

Déplacements de **bateaux**

Molina et al., 2011

Systèmes de contrôle des **eaux**

Molina et al., 2011

Informations **cliniques** pour le suivi de patients

Hüske-Kraus, 2003; Harris, 2008; Portet et al., 2009;

Gatt et al., 2009; Banaee et al., 2013

Monitoring de plateformes **industrielles**

Sripada et al., 2003; Yu et al., 2006

...

# Génération automatique de texte (NLG)

**text-to-text**



**data-to-text**

0	1	0	0
0	0	0	0
1	0	1	1



**image-to-text**



# Génération automatique à partir d'images

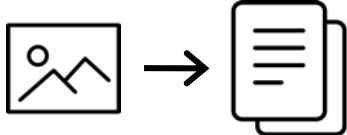
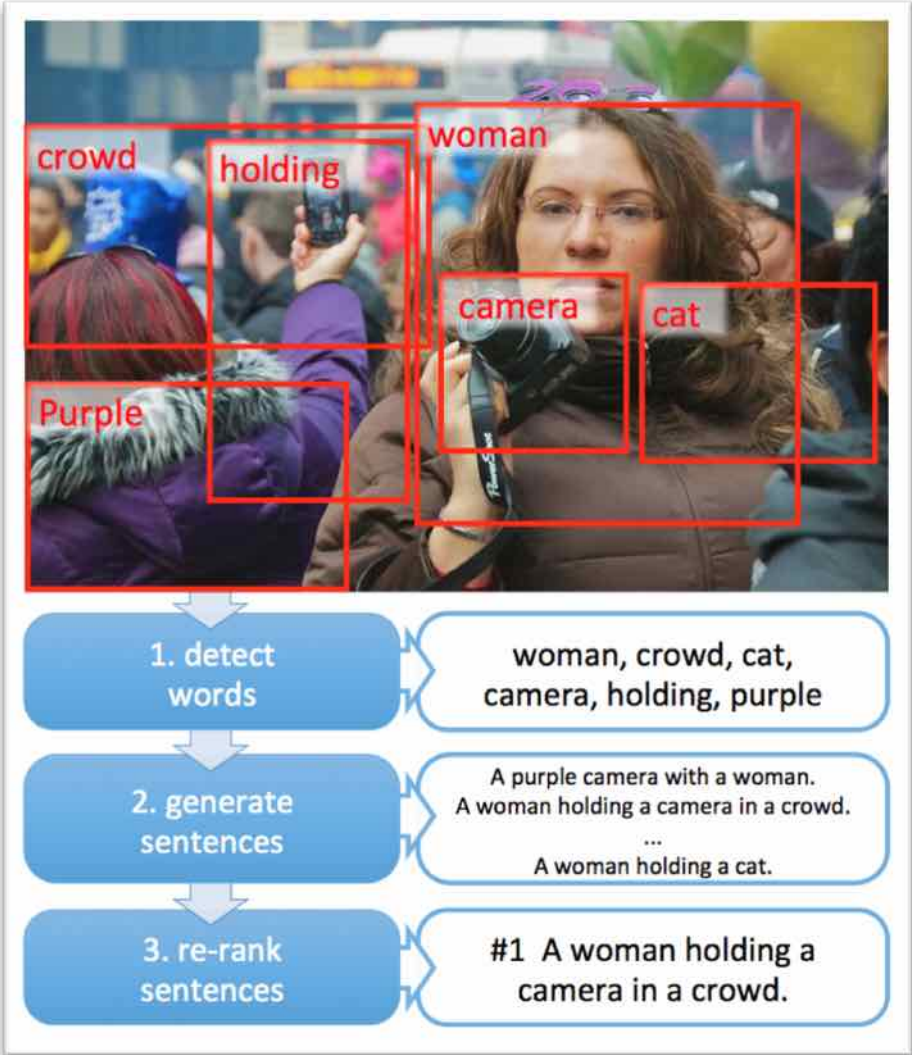


image-to-text

## Génération automatique de **légendes**

Platt, 2014 [Microsoft Cortana]

## Description automatique d'images pour **non-voyants**



# Génération automatique de texte (NLG)

**text-to-text**

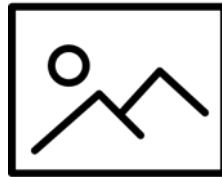


**data-to-text**

0	1	0	0
0	0	0	0
1	0	1	1



**image-to-text**



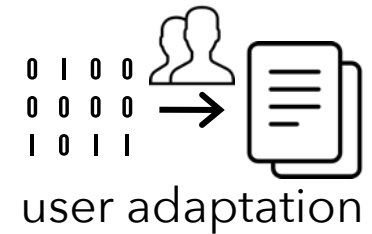
# Génération automatique de texte (NLG)

**adaptation**  
à l'utilisateur

0	1	0	0
0	0	0	0
1	0	1	1



# Adaptation visant l'accessibilité



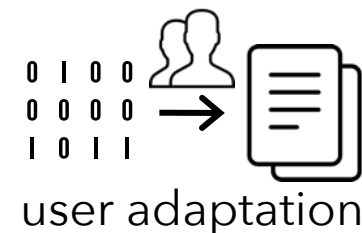
Description de **graphes** et de **schémas**  
pour non-voyants

## iGraph-Lite



The line graph has the title "operating profits up slightly in second quarter". Quarterly values from 1996 to 2006 are shown. The values are given in billions of dollars. The lowest yearly values occurred in the second quarter of 1996 (26.7 billion dollars), and the highest values in the fourth quarter of 2005 (57.6 billion dollars).

# Adaptation au **niveau de compétence**



## **Simplification** automatique de texte

Analyse de **lisibilité**

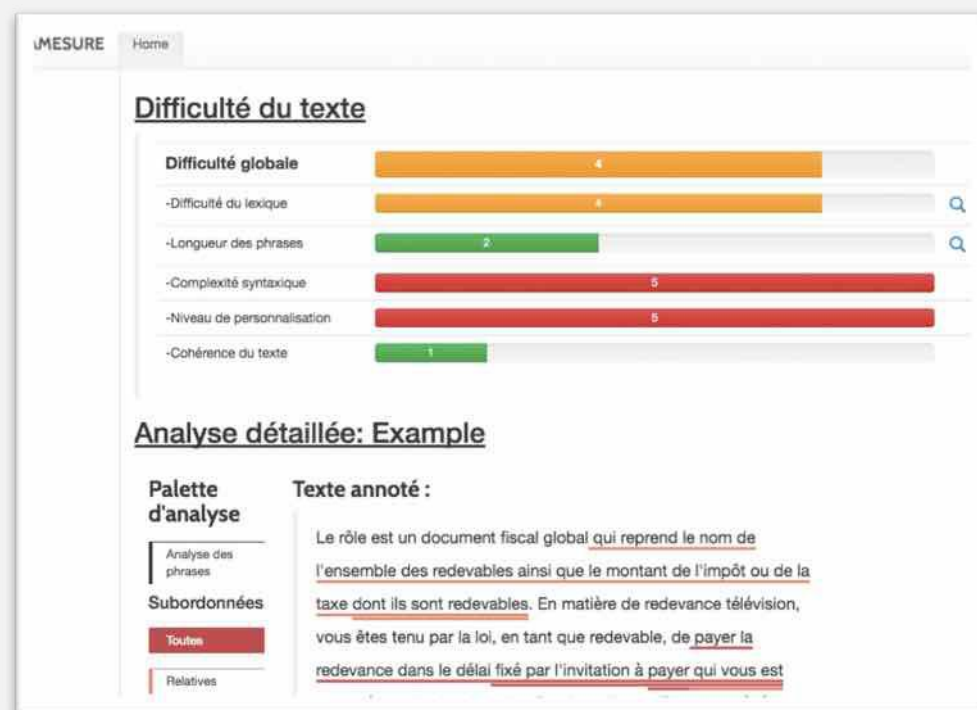
**aMesure**

(T. François, UCL/CENTAL)

**dMesure** → pour non-natifs

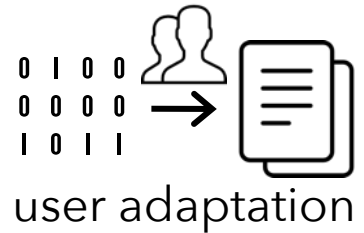
(T. François, UCL/CENTAL)

**Vocabulaire** plus ou moins précis selon le niveau de **spécialisation** du lecteur.





# Adaptation aux **sensibilités**

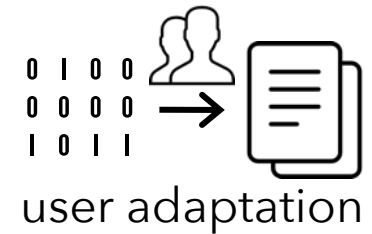


Adaptation aux **sensibilités politiques** ou aux profils **psychologiques** pour mieux convaincre ou être plus accessible

Description d'un match selon la **perspective des fans** de chaque équipe

« Anderlecht donne une leçon de football aux Liégeois » vs. « Le Standard redouble de malchance face aux mauves »

# Adaptation aux intérêts et besoins



Aller à l'essentiel vs. Dans les détails  
selon le profil du lecteur

**BabyTalk** : suivi clinique de prématurés ; rapports  
différents pour médecins, infirmières et parents  
(Mahamood & Reiter, 2011)

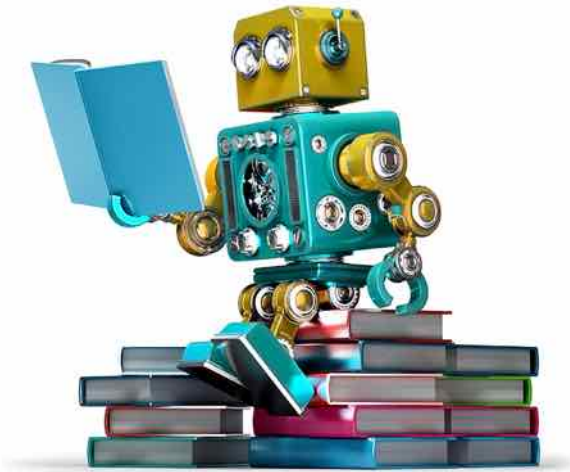


Currently, the baby is on CPAP in 27% O<sub>2</sub>. CPAP pressure is 4.4 cms H<sub>2</sub>O. SaO<sub>2</sub> is variable within the acceptable range and there have been

Since last week, his inspired Oxygen (FiO<sub>2</sub>) was lowered from 56% to 21% (which is the same as normal air). This is a positive development for your child.

# Génération automatique de texte (NLG)

## Approches du problème



### **Structuration modulaire classique**

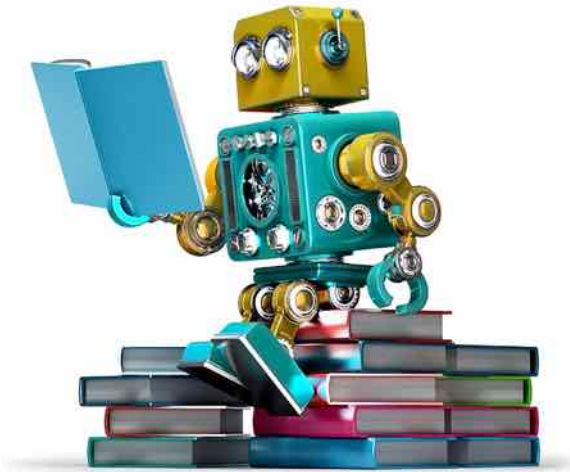
Rule-based, pipeline

### **Vers des approches probabilistes**

Data-driven, machine learning, deep learning

# Génération automatique de texte (NLG)

## Approches du problème



### **Structuration modulaire classique**

Rule-based, pipeline

### **Vers des approches probabilistes**

Data-driven, machine learning, deep learning

# Approche modulaire classique

## *NLG pipeline*



# Approche modulaire classique

## *Text planner* (Que dire ?)



Abstraction et interprétation  
des données

Détermination du contenu

Sélection des  
informations pertinentes

Structuration du texte

Détermination de la  
structure du message

# Approche modulaire classique

## *Sentence planner* (Comment le dire ?)



**Agrégation des phrases**

**Lexicalisation**

Mise de mots sur les "idées" à exprimer

**Génération des expressions référentielles**

Pour les expressions faisant référence à des éléments-clés ou récurrents du message, choix de termes fixes et discriminants

# Approche modulaire classique

## *Surface realizer* (En faire des phrases)



### Templates

\$joueur a marqué pour \$equipe à la \$minute-cardinal minute.

### Grammaires 'artisanales'

Plus flexible, mais très complexe à construire

$$\begin{aligned} S &\rightarrow NP VP \\ NP &\rightarrow D N \\ VP &\rightarrow V NP \end{aligned}$$


“Le Anderlecht bat par Standard 3-0”  
“Anderlecht bat le Standard par 3-0”

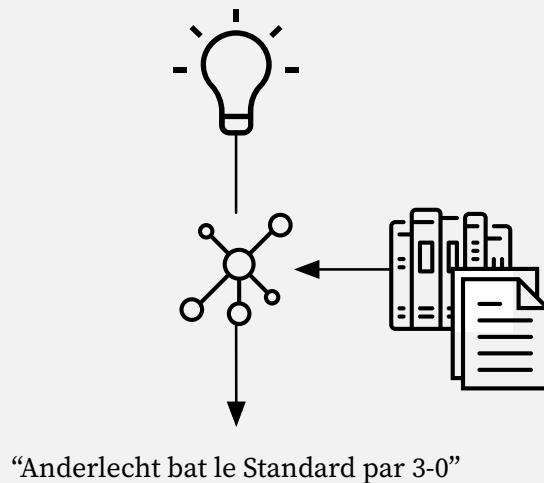
### Approches probabilistes (sur corpus)

Data-driven, statistical, machine learning

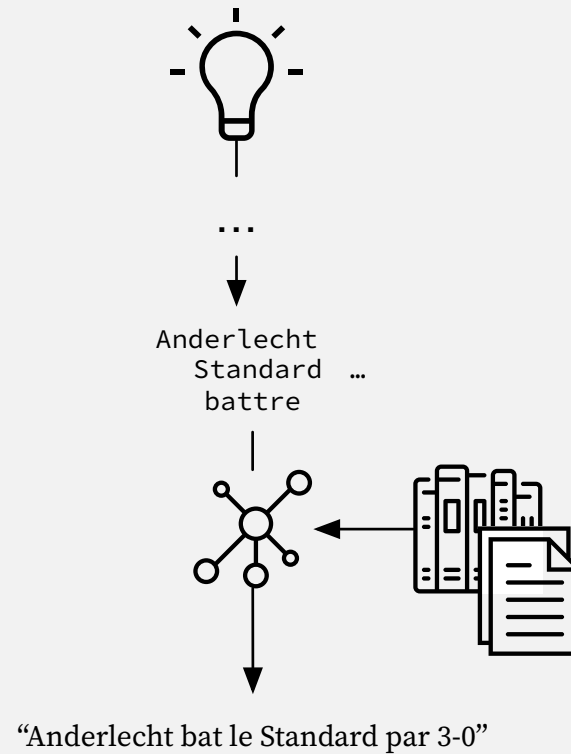


# Approche modulaire classique

## Réalisation linguistique probabiliste



**end-to-end** model

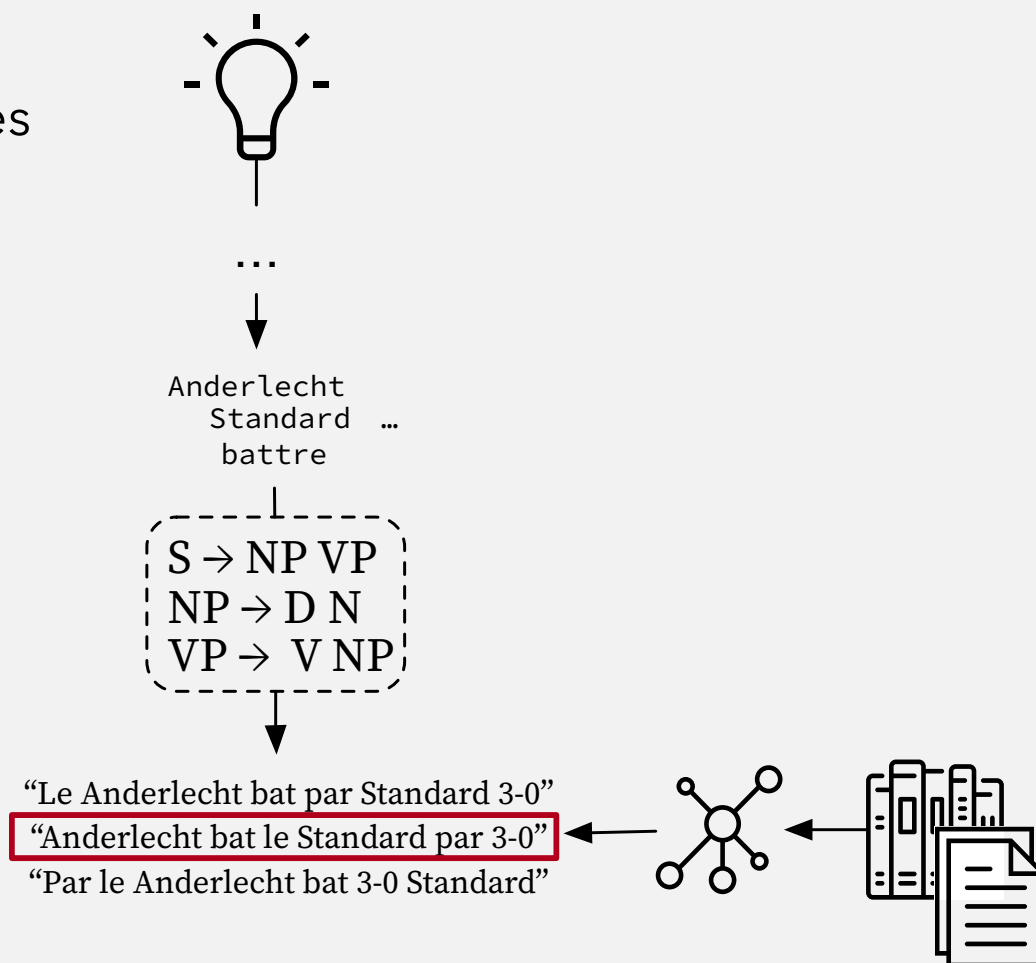


**realizer** model

# Approche modulaire classique

## Réalisation linguistique probabiliste

Génération 'artisanale' de candidats et **classement** des candidats sur corpus

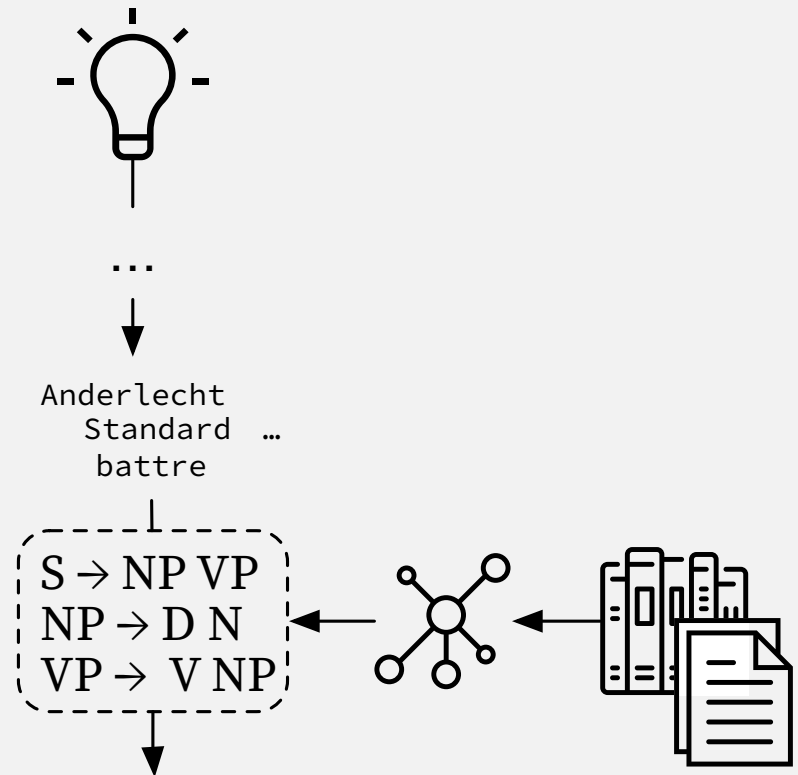


# Approche modulaire classique

## Réalisation linguistique probabiliste

Génération 'artisanale' de candidats et **classement** des candidats sur corpus

Génération 'artisanale' opérant des choix sur base d'un **apprentissage sur corpus**



“Anderlecht bat le Standard par 3-0”

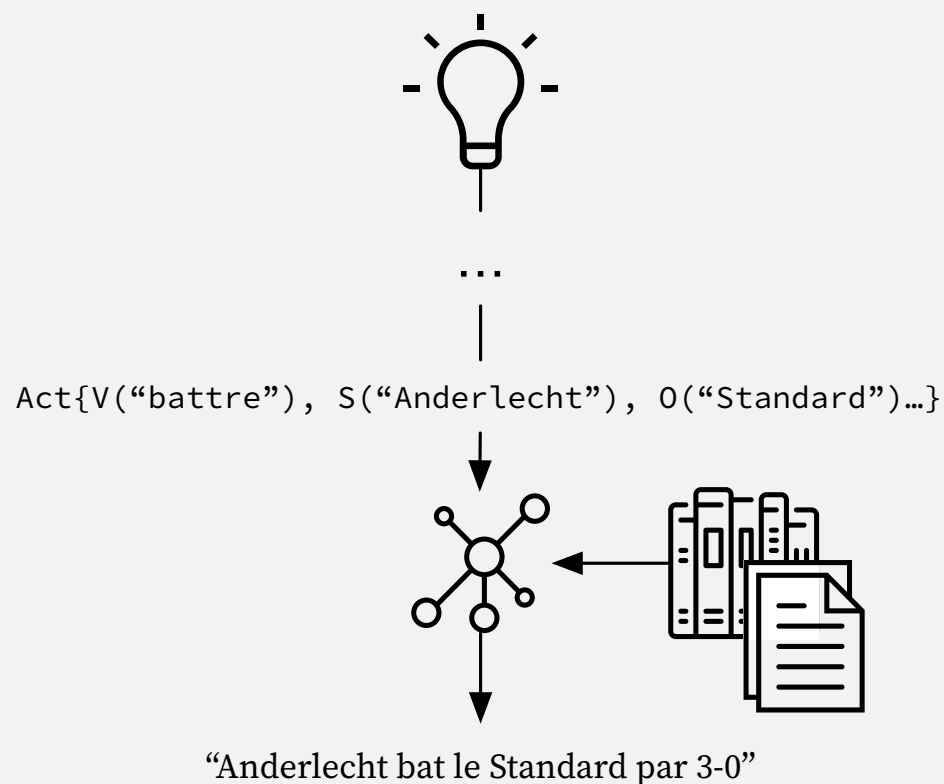
# Approche modulaire classique

## Réalisation linguistique probabiliste

Génération 'artisanale' de candidats et **classement** des candidats sur corpus

Génération 'artisanale' opérant des choix sur base d'un **apprentissage sur corpus**

Génération statistique (sur corpus) à partir d'un **formalisme grammatical**



# Approche modulaire classique

## Réalisation linguistique probabiliste

Génération 'artisanale' de candidats et **classement** des candidats sur corpus

Génération 'artisanale' opérant des choix sur base d'un **apprentissage sur corpus**

Génération statistique (sur corpus) à partir d'un **formalisme grammatical**

Combinatory Categorical Grammar (CCG)  
Steedman, 2000

Head-Driven Phrase Structure Grammar (HPSG)  
Nakanishi et al., 2005; Carroll & Oepen, 2005

Lexical-Functional Grammar (LFG)  
Cahill & Josef, 2006

Tree Adjoining Grammar (TAG)  
Gardent & Narayan, 2015

# Approche modulaire classique

## Limites

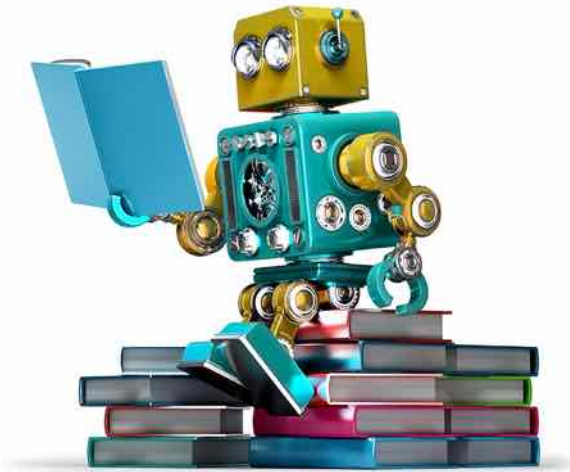
Étapes non cloisonnées,  
ou structurées différemment

Informations non disponibles avant la  
réalisation, mais nécessaires pour mieux  
informer les étapes précédentes  
(ex : contrainte de longueur du texte)

→ Tendances à intégrer les différentes étapes

# Génération automatique de texte (NLG)

## Approches du problème



### **Structuration modulaire classique**

Rule-based, pipeline

### **Vers des approches probabilistes**

Data-driven, machine learning, deep learning

# Approches probabilistes

Tendance générale (TAL, IA...): *data-driven approaches, big data, machine learning, deep learning...*

Entraîner un **modèle statistique/probabiliste** à associer un **input** à un **output** sur base d'un **corpus** (où les inputs et outputs sont déjà associés)

Au niveau d'une **sous-tâche** (réalisation...) ou pour l'ensemble de la génération ("**end-to-end**")

Résultats les plus convaincants obtenus sur base de corpus gigantesques (*big data*) et avec des réseaux de neurones complexes (*deep learning, neural networks*)



# Approches probabilistes

## Obtention des données d'entraînement

Besoin d'énormément de données **annotées** !  
(millions de correspondances  $I \rightarrow O$ )

Données "end-to-end" **disponibles** pour certains **domaines** du NLG

**Sport** : statistiques de matchs  $\Leftrightarrow$  résumés de matchs

**Finance** : statistiques boursières  $\Leftrightarrow$  dépêches financières

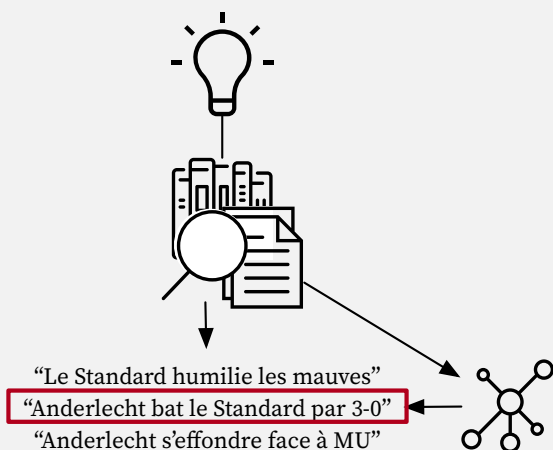
**Image-to-text** : photo  $\Leftrightarrow$  légende

**Météo** : données brutes météorologiques  $\Leftrightarrow$  annonces météos

Possible aussi d'utiliser le **crowdsourcing** pour annoter des grandes quantités de données existantes

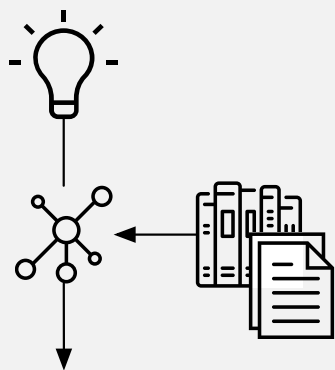
# Approches probabilistes

## *Retrieval vs. Generative*



**Approche 'retrieval':** récupération d'exemples existants du corpus d'entraînement, puis classement et sélection de la meilleure proposition

- seulement si nouveaux cas peuvent être décrits avec des phrases existantes
- + plus fiable



**Approche générative:** apprentissage des règles de construction de phrases et recréation d'une nouvelle phrase.

- + plus flexible, s'adapte à de nouveaux cas
- moins fiable : phrases agrammaticales...

# Approches probabilistes

## *Deep learning*

Majoritairement **text-to-text** (e.a. sequence-to-sequence)

notamment pour systèmes de dialogue, traduction automatique...

Quelques exemples de **data-to-text**

Long short-term memory (LSTM)  
sur données météo

Mei et al., 2016

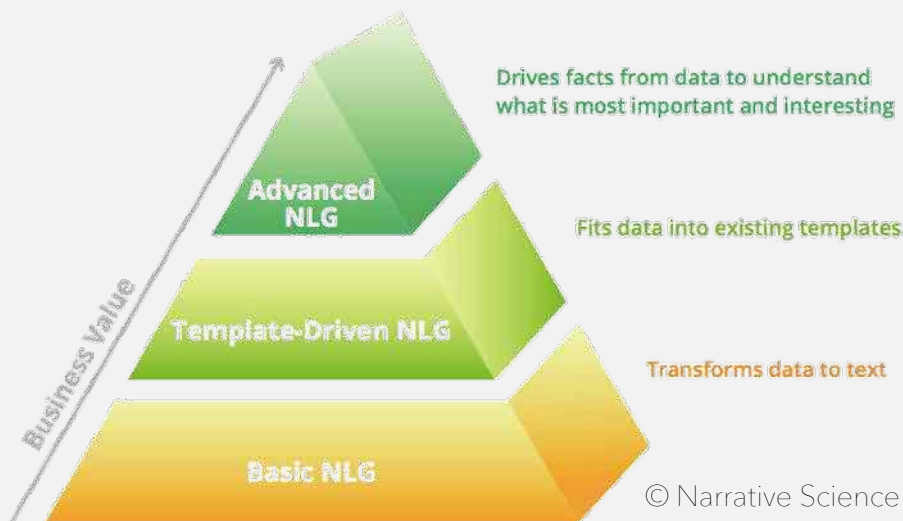
Seulement pour objectifs de recherche,  
pas en production

# Approches probabilistes

## Mais dans les applications finales...

Reste majoritairement sur des approches par **règles**, plus fiables

Revendications « avancées », « data-driven », « machine learning » des outils existants ?



→ IA oui, mais pour  
**l'interprétation / sélection / adaptation des données**,  
pas pour la réalisation linguistique

# Génération automatique de texte (NLG)

## En résumé...

Approches **probabilistes** prometteuses  
mais inutilisables en production

Trop d'erreurs, trop peu de contrôle

Techniques de *machine learning* (IA)

✓ pour la préparation des données

sélection, interprétation, adaptation

✗ pour la génération linguistique

→ Réalisation linguistique par *templating*

Applications novatrices de  
**personnalisation/adaptation** de  
l'information à l'utilisateur

```
NLGElement action = nlgFact.createVerbPhrase("remercier");  
action.setSubject(nlgFact.createNounPhrase("je"));  
action.setObject(nlgFact.createNounPhrase("vous"));  
action.addComplement(nlgFact.createPrepositionPhrase("pour votre attention");  
  
String output = realiser.realiseSentence(action);  
System.out.println(output);
```

**je vous remercie pour votre attention**



Serge Bibauw  
serge.bibauw@uclouvain.be

Téléchargez cette présentation sur  
<http://sbibauw.github.io/talk/audaxis>

# Références

- Gatt & Krahmer, 2017: *Survey of the state of the art in natural language generation*
- Reiter & Dale, 2000: *Building natural language generation systems*
- Arria NLG, *Technical Overview The Arria NLG Engine* (2015)
- Narrative Science, *The Automated Analyst: Transforming Data into Stories with Advanced Natural Language Generation*
- Marr, Bernard (2015). *Can Big Data Algorithms Tell Better Stories Than Humans?* in *Forbes*