Proposition d'architecture à base de corpus pour la Génération Automatique de Texte

Eric Charton

Séminaires du Rali, Montréal
Février 2010

Plan

La génération automatique de texte (GAT)

Objectif

Produire un texte en langue naturelle à partir d'une représentation formelle d'un contenu

La génération automatique de texte (GAT)

Objectif

Produire un texte en langue naturelle à partir d'une représentation formelle d'un contenu

Exemples applicatifs

La génération automatique de texte (GAT)

Objectif

Produire un texte en langue naturelle à partir d'une représentation formelle d'un contenu

Exemples applicatifs

- Produire un bulletin météo d'après des données [SumTime-Mousam - Sripada 03]

La génération automatique de texte (GAT)

Objectif

Produire un texte en langue naturelle à partir d'une représentation formelle d'un contenu

Exemples applicatifs

- Produire un bulletin météo d'après des données [SumTime-Mousam - Sripada 03]
- Réponse automatique à des E.Mails [Lapalme - 03]

La génération automatique de texte (GAT)

Objectif

Produire un texte en langue naturelle à partir d'une représentation formelle d'un contenu

Exemples applicatifs

- Produire un bulletin météo d'après des données [SumTime-Mousam - Sripada 03]
- Réponse automatique à des E.Mails [Lapalme - 03]
- Produire un texte d'après un formulaire [Smoking information questionnaire - Aberdeen NLG group 99]

La génération automatique de texte (GAT)

Objectif

Produire un texte en langue naturelle à partir d'une représentation formelle d'un contenu

Exemples applicatifs

- Produire un bulletin météo d'après des données [SumTime-Mousam - Sripada 03]
- Réponse automatique à des E.Mails [Lapalme - 03]
- Produire un texte d'après un formulaire [Smoking information questionnaire - Aberdeen NLG group 99]
- Documentation industrielle automatisée [Automatic Generation of Technical documentation - Reiter 95]

La génération automatique de texte (GAT)

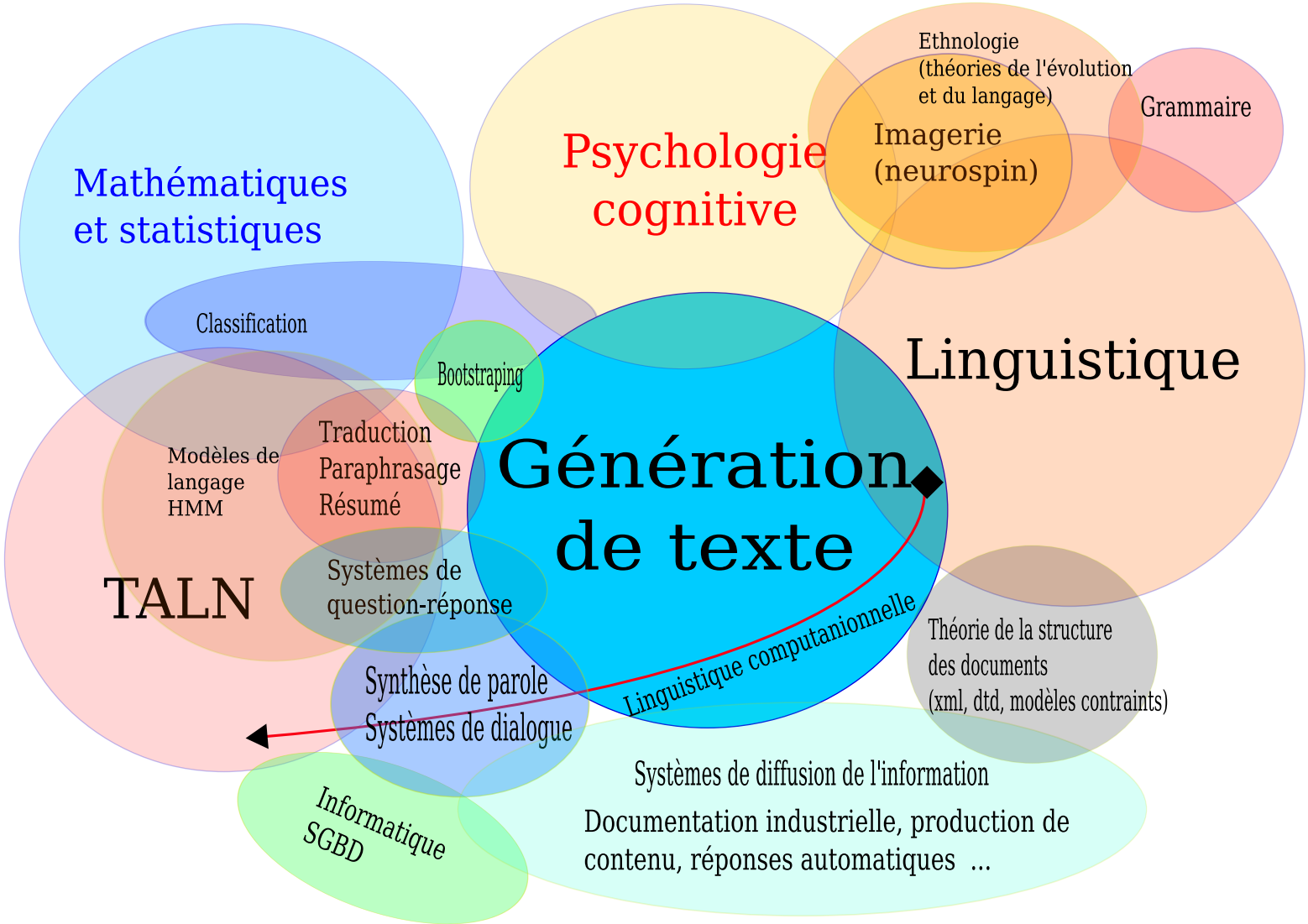
Objectif

Produire un texte en langue naturelle à partir d'une représentation formelle d'un contenu

Exemples applicatifs

- Produire un bulletin météo d'après des données [SumTime-Mousam - Sripada 03]
- Réponse automatique à des E.Mails [Lapalme - 03]
- Produire un texte d'après un formulaire [Smoking information questionnaire - Aberdeen NLG group 99]
- Documentation industrielle automatisée [Automatic Generation of Technical documentation - Reiter 95]
- Génération de réponses dans un système de dialogue [Rambow et al 01]

Un domaine de recherche largement pluridisciplinaire



Qu'est ce qu'un texte ?

Une hiérarchie

- Un document composé de paragraphes (le plan)

Qu'est ce qu'un texte ?

Une hiérarchie

- Un document composé de paragraphes (le plan)
- Des paragraphes composés de phrases (le contenu)

Qu'est ce qu'un texte ?

Une hiérarchie

- Un document composé de paragraphes (le plan)
- Des paragraphes composés de phrases (le contenu)
- Des phrases composées de mots (le style)

Qu'est ce qu'un texte ?

Une hiérarchie

- Un document composé de paragraphes (le plan)
- Des paragraphes composés de phrases (le contenu)
- Des phrases composées de mots (le style)

Générer du texte commence par la production de phrases : qu'est ce qu'une phrase ?

Qu'est ce qu'un texte ?

Une hiérarchie

- Un document composé de paragraphes (le plan)
- Des paragraphes composés de phrases (le contenu)
- Des phrases composées de mots (le style)

Générer du texte commence par la production de phrases : qu'est ce qu'une phrase ?

- Un système infini et non dénombrable, soumis à une grammaire transformationnelle et non modélisable par des approches statistiques (Chomsky)

Qu'est ce qu'un texte ?

Une hiérarchie

- Un document composé de paragraphes (le plan)
- Des paragraphes composés de phrases (le contenu)
- Des phrases composées de mots (le style)

Générer du texte commence par la production de phrases : qu'est ce qu'une phrase ?

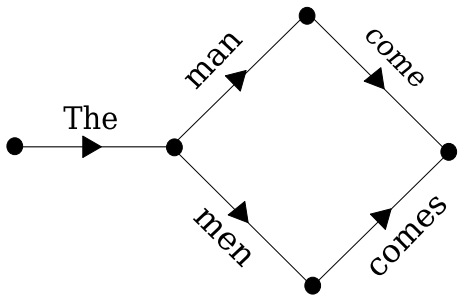
- Un système infini et non dénombrable, soumis à une grammaire transformationnelle et non modélisable par des approches statistiques (Chomsky)
- Chaque phrase est finie, l'ensemble des phrases est infini mais dénombrable. La langue est régie par la distribution des mots et groupes de mots (Harris)

Les phrases selon Chomsky

Chomsky

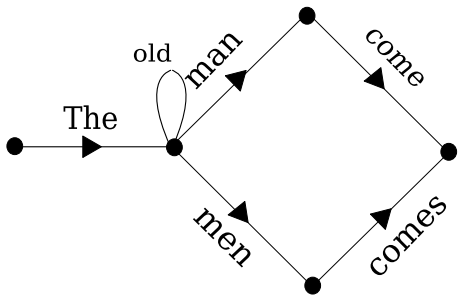
Pour Chomsky, les processus de Markov à nombre finis d'états ne permettent pas de modéliser une langue : *"Il est impossible de construire une machine qui produirait [toutes] les phrases grammaticales de l'Anglais"*

Chomsky : structures syntaxiques



The man come

The mens comes



The old man come

The old men comes

...

The old old men comes

...

Les phrases selon Harris

Harris : analyse distributionnelle

- "On peut décrire toute langue par une structure distributionnelle, cad, par l'occurrence des parties relativement les unes aux autres"

A = e + f et B = e + g
A dérive de B par substitution de g à f
Exemple : il est [en outre] très poli -> il est [par ailleurs] très poli

Les phrases selon Harris

Harris : analyse distributionnelle

- "On peut décrire toute langue par une structure distributionnelle, cad, par l'occurrence des parties relativement les unes aux autres"
- "[Le modèle distributionnel] consiste à décrire toutes les formes [linguistiques] comme des combinaisons d'éléments"

A = e + f et B = e + g
A dérive de B par substitution de g à f
Exemple : il est [en outre] très poli -> il est [par ailleurs] très poli

Les phrases selon Harris

Harris : analyse distributionnelle

- "On peut décrire toute langue par une structure distributionnelle, cad, par l'occurrence des parties relativement les unes aux autres"
- "[Le modèle distributionnel] consiste à décrire toutes les formes [linguistiques] comme des combinaisons d'éléments"
- "Une forme A est dérivée d'une forme B par substitution [de ses éléments]"

A = e + f et B = e + g
A dérive de B par substitution de g à f
Exemple : il est [en outre] très poli -> il est [par ailleurs] très poli

Les phrases selon Shannon

Génération de phrases selon un modèle de language

- Dans *A mathematical theory of communication*

THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED.

Les phrases selon Shannon

Génération de phrases selon un modèle de language

- Dans *A mathematical theory of communication*
- *"The resemblance to ordinary English text increases quite noticeably at each of the above steps"*

THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED.

Evolution de l'architecture des systèmes de génération automatique de texte

Une mise au point progressive depuis les années 50

Approches statistiques et combinatoires

- Tentative d'implémenter les grammaires dans des systèmes de génération combinatoires [Mathews,1962]

Une mise au point progressive depuis les années 50

Approches statistiques et combinatoires

- Tentative d'implémenter les grammaires dans des systèmes de génération combinatoires [Mathews,1962]
- Référence explicite à la théorie de la communication de Shannon

Une mise au point progressive depuis les années 50

Approches statistiques et combinatoires

- Tentative d'implémenter les grammaires dans des systèmes de génération combinatoires [Mathews,1962]
- Référence explicite à la théorie de la communication de Shannon

L'influence Chomskyennes, années 60

Une mise au point progressive depuis les années 50

Approches statistiques et combinatoires

- Tentative d'implémenter les grammaires dans des systèmes de génération combinatoires [Mathews,1962]
- Référence explicite à la théorie de la communication de Shannon

L'influence Chomskyennes, années 60

- Génération de phrases par combinaisons et introduction de règles [Yngve,1960]

Une mise au point progressive depuis les années 50

Approches statistiques et combinatoires

- Tentative d'implémenter les grammaires dans des systèmes de génération combinatoires [Mathews,1962]
- Référence explicite à la théorie de la communication de Shannon

L'influence Chomskyennes, années 60

- Génération de phrases par combinaisons et introduction de règles [Yngve,1960]
- Génération à base exclusive de grammaires hors contexte [Friedman, 1969]

Années 70 à nos jours, 3 propositions dominantes

Approches par patrons à trous

- Des modèles de phrases prédéfinis contenant des éléments variables [Reiter 1995 ;Deemter 2005]

Années 70 à nos jours, 3 propositions dominantes

Approches par patrons à trous

- Des modèles de phrases prédéfinis contenant des éléments variables [Reiter 1995 ;Deemter 2005]

Approches à base de règles

- systèmes à base de règle et de grammaires régis par une architecture modulaire en pipeline [Reiter, 1994 ; Lapalme & Danlos, 2000]

Années 70 à nos jours, 3 propositions dominantes

Approches par patrons à trous

- Des modèles de phrases prédéfinis contenant des éléments variables [Reiter 1995 ;Deemter 2005]

Approches à base de règles

- systèmes à base de règle et de grammaires régis par une architecture modulaire en pipeline [Reiter, 1994 ; Lapalme & Danlos, 2000]

Approches statistiques et n-grammes

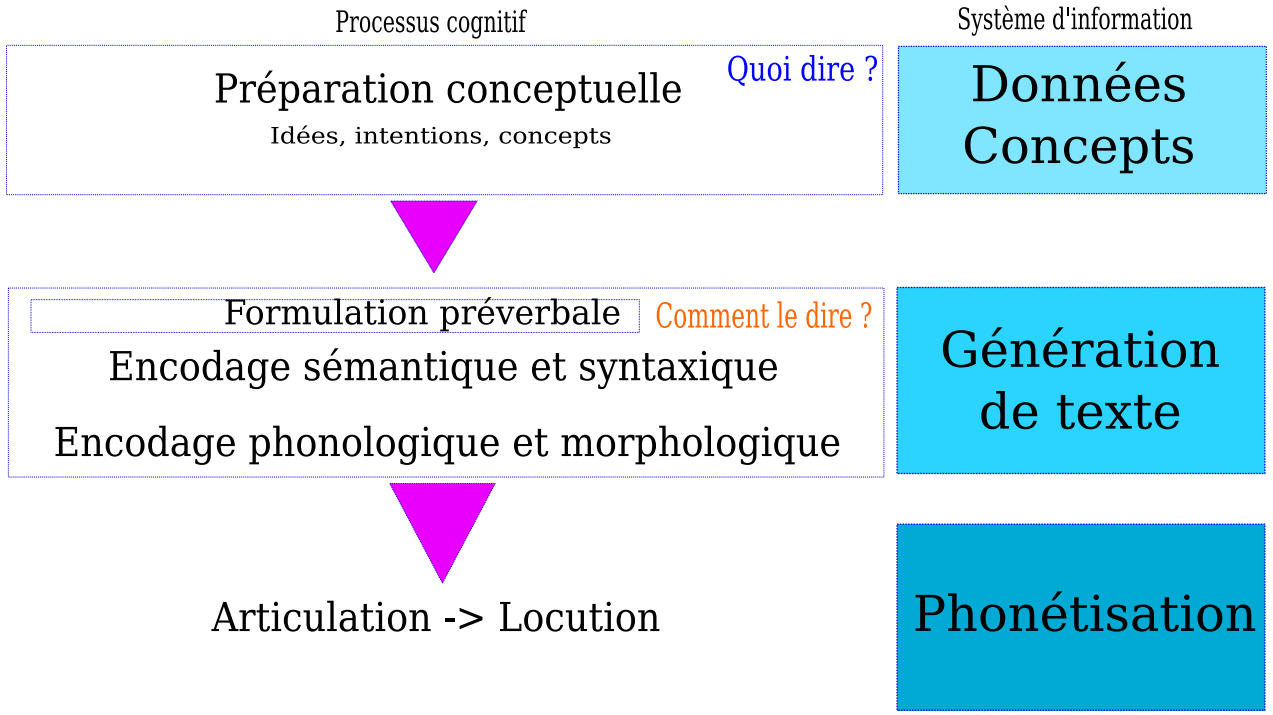
- systèmes probabilistes guidés reposant sur des assemblages de n-grammes [Langdike & Knight, 1998 ; Belz, 2006]

Fonctionnement des systèmes de génération automatique de texte

Que dire et comment le dire ?

Deux paradigmes incontournables

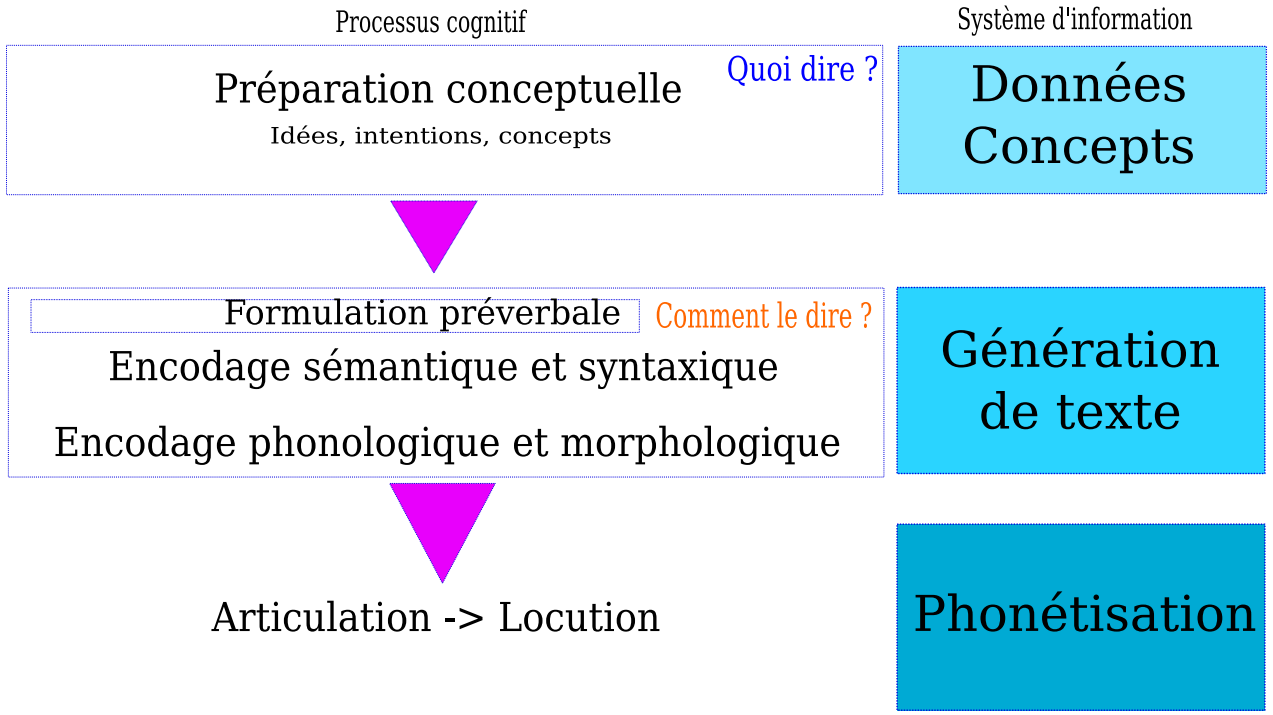
- "Quoi dire" et "Comment le dire"



Que dire et comment le dire ?

Deux paradigmes incontournables

- "Quoi dire" et "Comment le dire"
- Une caution psycholinguistique [Levelt, 89 ; Ferrand 02]



Le modèle de patron à trous

Un modèle classique et simple

- Le système de [Buseman, 1998] pour produire des bulletins de pollution
- Facile à déployer, aisé à maintenir en plusieurs langues
- Une hybridation simple avec les modèles plus complexes (voir comparaison par [Deemter 2005])

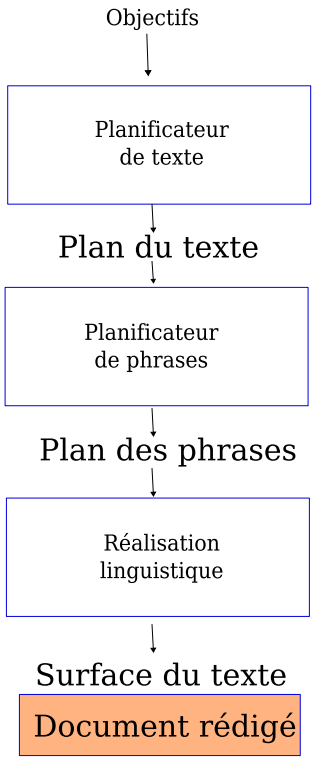
```
[(COOP THRESHOLD-EXCEEDING)
 (LANGUAGE FRENCH)
 (TIME [(PRED SEASON) (NAME [(SEASON WINTER) (YEAR 1996)])])
 (THRESHOLD-VALUE [(AMOUNT 600) (UNIT MKG-M3)])
 (POLLUTANT SULFUR-DIOXIDE)
 (SITE "V&o1lklingen-City")
 (SOURCE [(LAW-NAME SMOGVERORDNUNG) (THRESHOLD-TYPE VORWARNSTUFE)])
 (DURATION [(HOUR 3)])
 (EXCEEDS [(STATUS NO) (TIMES 0)])]
```

En hiver 1996/97 à la station de mesure de Völklingen-City, le seuil d'avertissement pour le dioxide de soufre pour une exposition de trois heures (600.0 $\mu g/m^3$ selon le decret allemand "Smogverordnung") n'a pas été dépassée.

Le modèle d'architecture générique

Pipelined Natural langage Generation System

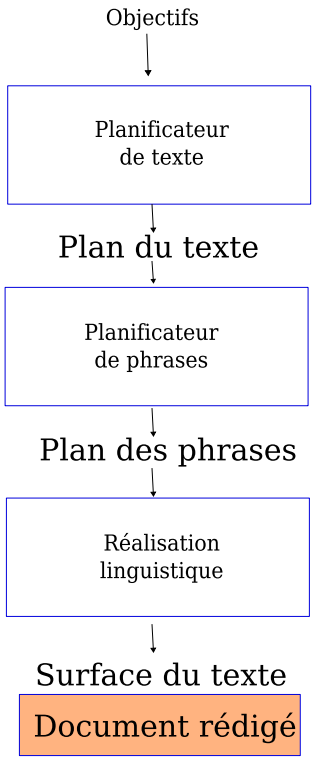
- Un ensemble de modules consécutifs



Le modèle d'architecture générique

Pipelined Natural langage Generation System

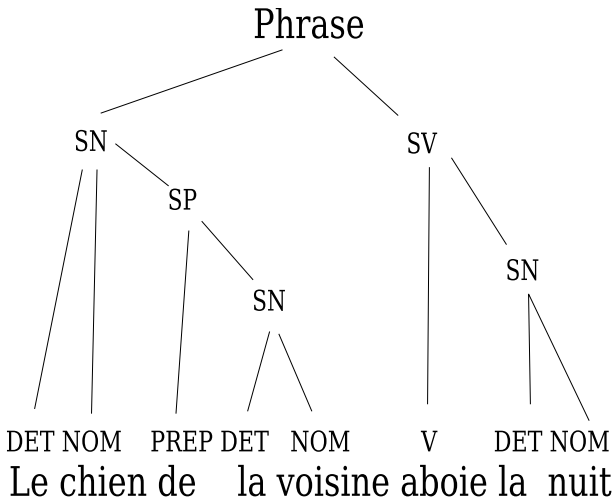
- Un ensemble de modules consécutifs
- Repose essentiellement sur des modèles formels



Production des phrases

Utilisation de modèles formels de représentation

- Arbres syntaxiques

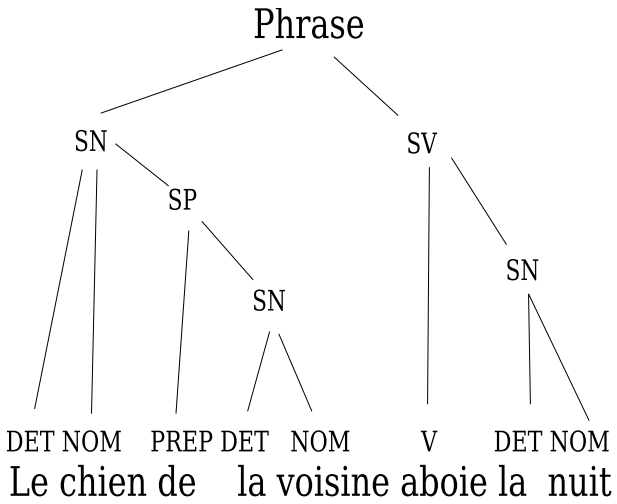



```
[ouvrir
    [case frame
        agent: voisine
        objet : chien
        action: aboyer
    [mode
         temps:présent
    ]
```

Production des phrases

Utilisation de modèles formels de représentation

- Arbres syntaxiques
- Réseaux de transitions

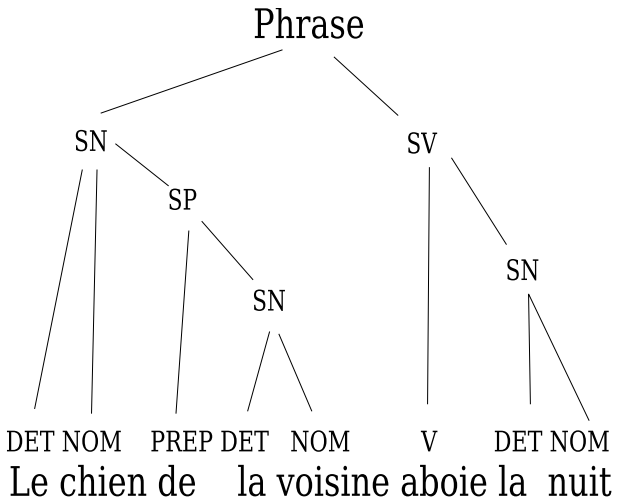



```
[ouvrir
    [case frame
        agent: voisine
        objet : chien
        action: aboyer
    [mode
         temps:présent
    ]
```

Production des phrases

Utilisation de modèles formels de représentation

- Arbres syntaxiques
- Réseaux de transitions
- Frames

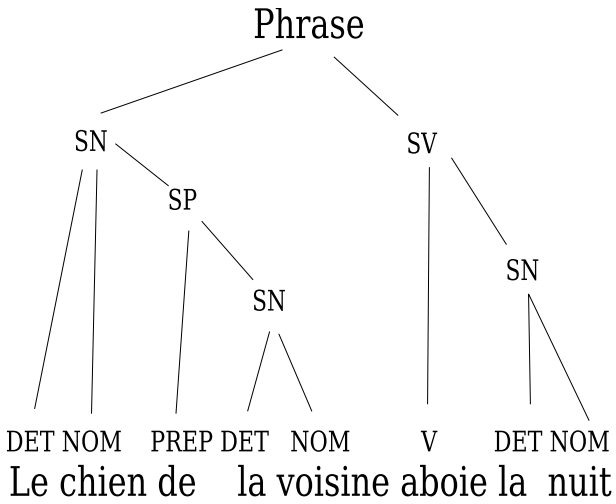



```
[ouvrir
    [case frame
        agent: voisine
        objet : chien
        action: aboyer
    [mode
         temps:présent
    ]
```

Production des phrases

Utilisation de modèles formels de représentation

- Arbres syntaxiques
- Réseaux de transitions
- Frames
- Graphes conceptuels




```
[ouvrir
    [case frame
        agent: voisine
        objet : chien
        action: aboyer
    [mode
         temps:présent
    ]
```

Exemple de système à base de grammaires

Simple NLG [Reiter, 2009]

- Librairie en Java. Il faut *programmer* le texte : indiquer le temps, l'intention de communication
- Le système gère la construction de phrase, les conjugaisons, les connexions logiques
- Uniquement en anglais, très difficile à adapter

```
SPhraseSpec s1 = new SPhraseSpec("I", "be", "happy");
SPhraseSpec s2 = new SPhraseSpec("I", "eat", "fish");
s2.setCuePhrase("because");
s2.setTense(Tense.PAST);

TextSpec t1 = new TextSpec(); // create a TextSpec
t1.addSpec(s1);
t1.addSpec(s2);

String output = r.realiseDocument(t1);  //Realiser r created earlier
System.out.println(output);
```

The output is:

```
I am happy, because I ate fish.
```

Expériences à base de corpus

Le corpus en tant que ressource pour la génération

- Le corpus est utilisé en tant que ressource de n-grammes [Langkilde, 1998]
- Le corpus est utilisé en tant que ressource de choix lexical [Bangalore, 2000]
- Le corpus est utilisé en tant que ressource de parties de phrases avec des étiquettes discursives [Marciniak, 2005]
- Aucun système n'utilise le corpus en tant que ressource de phrases prêtes à l'emploi

Un système de génération automatique de texte d'après une librairie de phrases modèles

Principe de base

Tirer partie des avantages de chaque modèle

- Remplacer la génération syntaxique par un inventaire de phrases le plus grand possible
- Avantage : génération multilingue possible, adaptation des modèles de génération automatisée
- Utiliser le principe des patrons à trous pour transformer les contenus d'une phrase existante
- Avantage : simplicité du processus de transformation
- Utiliser des modèles n-grammes pour réaliser les dernières transformations de surface (genre, etc)
- Avantage : les phrases du modèle ont une meilleure couverture puisqu'elles deviennent partiellement transformables

Modélisation des phrases d'une langue (1)

Le modèle de phrases

- Récupérer depuis un corpus proche du domaine de génération plusieurs millions de phrases
- Le corpus ne sert plus à modéliser les n-grammes mais les formes de phrases
- Etiqueter ces phrases à plusieurs niveaux (lexical, morphosyntaxique, syntagmatique) pour les rendre abstraites

Les corpus de phrases possibles

- Wikipédia (28 millions de phrases FR, 115 millions de phrases EN ...)
- Wikisource, Gutemberg (plusieurs milliers de livres thématiques)
- Des corpus adaptés au domaine de génération visé (notions de e.langage et i.langage) (juridique, technique, web) (Notion de I-Language [Chomsky 1986] / masse parlante [Saussure 1894])

Modélisation des phrases d'une langue (2)

Le modèle de génération

- Formaliser une intention de communication : l'arbre de dépendance, les contenus des syntagmes, les concepts et leurs synonymes
- Mesurer la similarité entre l'intention de communication et les phrases abstraites contenues dans le modèle
- Sélectionner une liste des N meilleures phrases candidates
- Choisir la meilleure candidate et remplacer les abstractions par les éléments de l'intention de communication

Représentation d'une espace linguistique

Comment le dire ?

Le modèle de phrase à trois niveaux

- A : Niveau lexical et conceptuel
- B : Niveau morpho-syntaxique
- C : Niveau syntagmatique (dépendances)

Rendre les phrase du corpus les plus abstraites possibles

Exemple : **Le Rhône passe en bordure d'Avignon**

A	Le	LOC.GEO	passe	en	bordure	d'	LOC.ADMI
B	det.art	nam	verb.pres	prp	nom	prp	nam
C	SN	SN	NV	SA	SA	SA	SA

Représentation d'une intention de communication

Que dire ?

La représentation de l'intention de communication

- A : Concept lexical (réseau de synonynes) - Entités nommées
- B : Niveau morpho-syntaxique
- C : Niveau syntagmatique (dépendances)

Exemple : **Loire ;Couler :présent ;Autour ;Orléans**

LOC.GEO nam SN	couler :passer :ruisseler :traverser verb.pres NV	autour :bord :orée :pourtour :corniche nom SA	LOC.ADMI nam SA

Mesures de similarité

Chercher une phrase dans le modèle (M) correspondant à la représentation de l'intention de communication (I)

- 1 : évaluer le degré de proximité lexicale
- 2 : évaluer la proximité des arbres de dépendances
- 3 : évaluer la compatibilité de temps, de forme (négations, pluriels, etc)

Méthode utilisée

- 1 : La similarité cosinus permet de mesurer la proximité lexicale - $cos(M_{lex}, I_{lex})$
- 2 : un calcul de pourcentage de proximité p_s appliqué sur chaque niveau des arbres de M et I comparés
- 3 : Calcul de similarité cosinus sur les étiquettes de *POS* $cos(M_{pos}, I_{pos})$ (ie : temps des verbes)
- Rang de la **phrase candidate** : est égal au produit de $cos(M_{lex}, I_{lex}) * p_s * cos(M_{pos}, I_{pos})$ (ou à la somme des log base 10)

Appliquer un traitement de surface à la phrase choisie

I=Loire ;Couler⟶ présent ;Autour ;Orléans

Exemple : **M=Le Rhône passe en bordure d'Avignon**

A	Le	LOC.GEO	passe	en	bordure	d'	LOC.ADMI
B	det.art	nam	verb.pres	prp	nom	prp	nam
C	SN	SN	NV	SA	SA	SA	SA

Remplacer les contenus par les correspondances (principe du patron à trous)

A	Le	**Loire**	passe	en	bordure	d'	**Orléans**
B	det.art	nam	verb.pres	prp	nom	prp	nam
C	SN	SN	NV	SA	SA	SA	SA

Appliquer un traitement de surface avec des règles ou des modèles n-grammes

La Loire passe en bordure d' Orléans

Composants du système

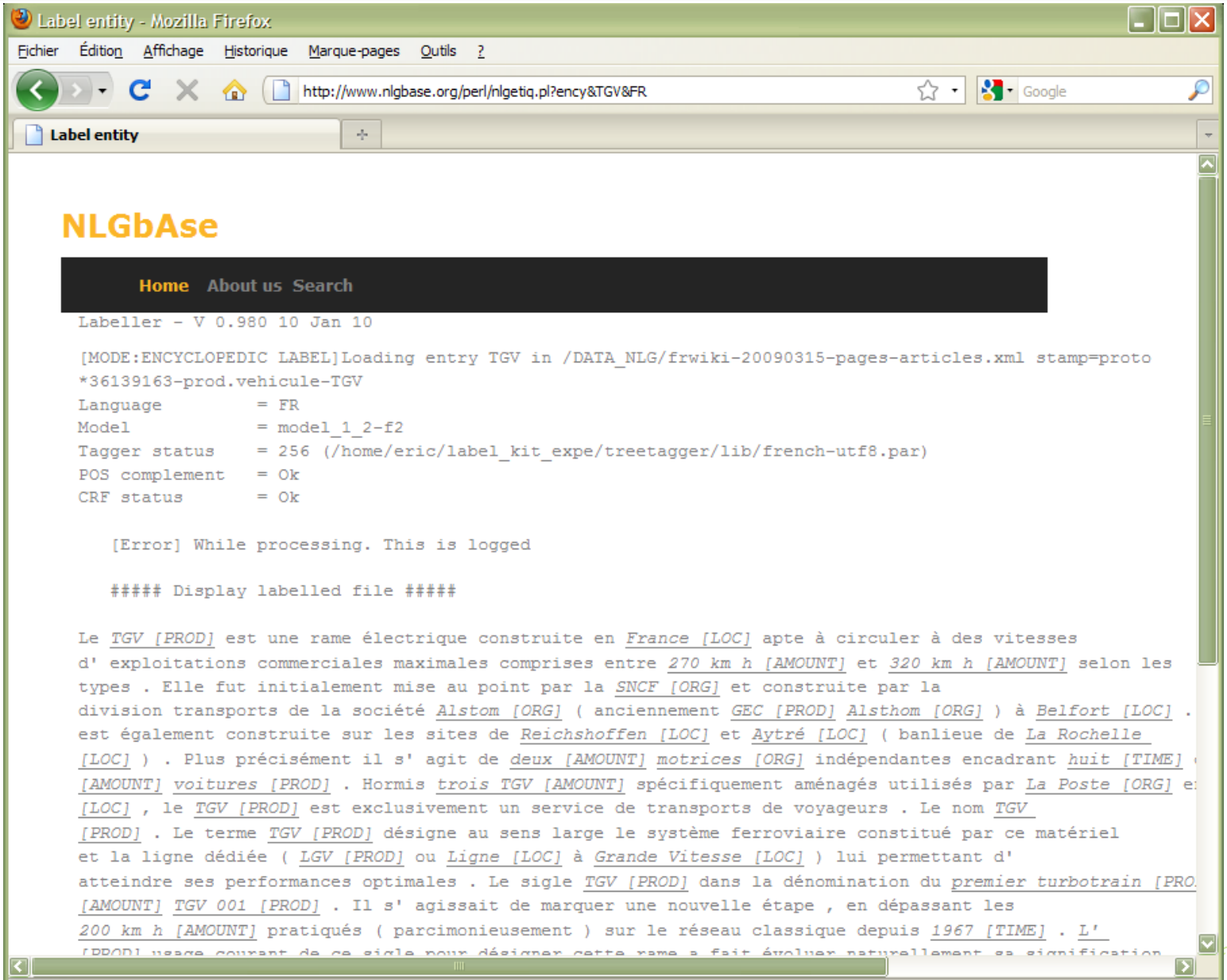
Génération du modèle de phrases

- Lexique d'entités nommées NLGbAse (5 langues) [Charton & Torres-Morenno, 2009]
- Etiqueteur d'entités nommées LIA/ESTER (CRF) [Béchet & Charton, 2010]
- Lexique terminologique et verbal Worldnet / base de synonymes Cortex
- Etiqueteur morphosyntaxique multilingue LIA-TAG/TreeTagger
- Algorithme de substitution par n-grammes [Charton & Torres-Morenno, 2010]

Modèles de phrases produits

- Corpus Wikipédia FR, EN, ES, IT, PL
- Modèle français : 28 millions de phrases
- Modèle anglais : 120 millions de phrases
- Modèle espagnol : 12 millions de phrases (en cours)

Composants du système



Evaluation et résultats préliminaires

Expérience

Principe

- Générer un modèle de phrase appris sur Wikipédia FR
- Extraire 100 phrases qui seront retirées du modèle
- Construire une représentation de l'*intention de communication* pour les 100 phrases
- Chercher des modèles de phrases compatibles
- Produire une phrase syntaxiquement et sémantiquement correcte

Expérience

Exemple

- Phrase servant de base pour construire l'intention de communication
- **Les armées de Junot envahissent le pays**

Formalisation

A	armée ;troupes ;	PERS	envahir attaquer	pays ;voisin ;frontière
B	nom	nam	verb.pres	nom
C	SN	SN	NV	SA

Expérience

Formalisation

A	armée ;troupes ;	PERS	envahir attaquer	pays ;voisin ;frontière
B	nom	nam	verb.pres	nom
C	SN	SN	NV	SA

Propositions ordonnée fournie par le système

- **Belka disposant d'une puissante armée , envahit son voisin**
- Les armées Wisigoth envahirent le pays
- Les princes ruthénes envahirent le pays polonais
- Le sire Anselme de Ribeaupierre attaqua en 1287 la ville de Saint-Hippolyte
- Les Philistins envahirent une fois de plus le pays
- Lesdites hostilités débutent lorsque les premiers attaquent à l'arme lourde le domicile privée de l'ex Président Sassou

Transformation

- **Junot** disposant d'une puissante **armée**, **envahit** son **voisin**

Résultats préliminaires

Validation de l'algorithme de similarité

100% des phrases qui correspondent à l'intention de communication sont retrouvées dans le modèle de phrases si elles y sont présentes

Validation du processus de génération pour 100 *intentions de communication*, non présentes dans le modèle de phrases

Sens et syntaxe correcte	74
Sens correct et syntaxe erronée	9
Sens incorrect et syntaxe correcte	6
Sens incorrect et syntaxe erronée	11

Conclusions

Une architecture prometteuse

- Le système fonctionne dans 74% des cas
- Il est moins performant qu'un système à base de règles et de grammaires qui fonctionne dans tous les cas
- Il est peu coûteux à mettre au point, facilement adaptable à plusieurs langues
- La taille du corpus d'apprentissage et son domaine jouent un rôle important sur la qualité et les performances

Perspectives

- Achever la mise au point (traitement des négations, etc)
- Produire un jeu d'expérience de taille suffisante
- Identifier une méthode d'évaluation semi-automatique
- Appliquer dans un contexte multilingue

Merci

Eric Charton - eric.charton@polymtl.ca