Compression automatique ou semi-automatique de textes par élagage des constituants effaçables : une approche interactive et indépendante des corpus

Mehdi Yousfi-Monod

LIRMM — Université Montpellier 2 — CNRS

Vendredi 16 novembre 2007

Présentation du jury

Compression automatique ou semi-automatique de textes par élagage des constituants effaçables : une approche interactive et indépendante des corpus

Thèse soutenue devant le jury suivant

Jacques Vergne	PR	Université de Caen	rapporteur
Jean-Luc Minel	IGR/HDR	Université Paris 10	rapporteur
Juan Manuel Torres-Moreno	MCF	Université d'Avignon	examinateur
Augusta Mela	MCF	Université Montpellier 3	examinateur
Jacques Chauché	PR	Université Montpellier 2	examinateur
Violaine Prince	PR	Université Montpellier 2	directrice de thèse

Contexte

Traitement Automatique des Langues Naturelles

Domaine d'étude des techniques automatiques d'analyse (compréhension) et de génération (production) d'énoncés oraux ou écrits.

Contexte

Traitement Automatique des Langues Naturelles

Domaine d'étude des techniques automatiques d'analyse (compréhension) et de génération (production) d'énoncés oraux ou écrits.

- la traduction automatique
- le résumé automatique
- la recherche d'information
- la correction orthographique
- la reconnaissance ou synthèse vocale
- la reconnaissance de l'écriture manuscrite
- ...

Contexte

Traitement Automatique des Langues Naturelles

Domaine d'étude des techniques automatiques d'analyse (compréhension) et de génération (production) d'énoncés oraux ou écrits.

- la traduction automatique
- le résumé automatique
- la recherche d'information
- la correction orthographique
- la reconnaissance ou synthèse vocale
- la reconnaissance de l'écriture manuscrite
- ...

Le résumé automatique

Le résumé automatique, principe général

À partir d'un document, produire par une méthode automatique un document **plus petit** qui soit **un bon représentant** de l'original.

Le résumé automatique

Le résumé automatique, principe général

À partir d'un document, produire par une méthode automatique un document **plus petit** qui soit **un bon représentant** de l'original.

Les styles des résumés :

- informatif : couvre l'ensemble des thèmes
- indicatif : propose un aperçu des thèmes
- agrégatif : fournit des informations non présentes ou non explicites
- critique : fournit une information critique sur le contenu

Le résumé automatique

Le résumé automatique, principe général

À partir d'un document, produire par une méthode automatique un document **plus petit** qui soit **un bon représentant** de l'original.

Les styles des résumés :

- informatif : couvre l'ensemble des thèmes
- indicatif : propose un aperçu des thèmes
- agrégatif : fournit des informations non présentes ou non explicites
- critique : fournit une information critique sur le contenu

Le résumé automatique

Le résumé automatique, principe général

À partir d'un document, produire par une méthode automatique un document **plus petit** qui soit **un bon représentant** de l'original.

Les styles des résumés :

- informatif : couvre l'ensemble des thèmes
- indicatif : propose un aperçu des thèmes
- agrégatif : fournit des informations non présentes ou non explicites
- critique : fournit une information critique sur le contenu

Objectifs :

- localiser les informations importantes et/ou peu importantes
- produire un document plus petit et **cohérent**

Mode de production

Résumé au sens humain

- travail de reformulation
- ✗ très difficile pour la machine
- ✗ peu d'approches, peu efficaces

Mode de production

Résumé au sens humain

- travail de reformulation
- ✗ très difficile pour la machine
- ✗ peu d'approches, peu efficaces

Résumé par extraction

- composition du résumé avec des fragments du document original
  $\Rightarrow$ supprimer des phrases et/ou réduire la taille des phrases
- ✗ vocabulaire limité à celui du document original
- ✓ abordable pour la machine

Mode de production

Résumé au sens humain

- travail de reformulation
- ✗ très difficile pour la machine
- ✗ peu d'approches, peu efficaces

Résumé par extraction

- composition du résumé avec des fragments du document original
  $\Rightarrow$ supprimer des phrases et/ou réduire la taille des phrases
- ✗ vocabulaire limité à celui du document original
- ✓ abordable pour la machine

Idée directrice de notre approche

Une phrase contient souvent de l'information peu importante

Idée directrice de notre approche

Une phrase contient souvent de l'information peu importante

Question : peut-on compresser un texte en supprimant l'information moins importante des phrases ?

Idée directrice de notre approche

Une phrase contient souvent de l'information peu importante

Question : peut-on compresser un texte en supprimant l'information moins importante des phrases ?

Exemple, extrait d'un article de « Le Monde.fr » (31/10/07)

Pour la première fois depuis les affrontements qui les ont opposés à la junte militaire à la fin du mois de septembre, entre cent et deux cents bonzes ont défilé pacifiquement pendant plus d'une heure dans la ville de Pakko, au centre de la Birmanie, mercredi 31 novembre, selon des témoins.

Idée directrice de notre approche

Une phrase contient souvent de l'information peu importante

Question : peut-on compresser un texte en supprimant l'information moins importante des phrases ?

Exemple, extrait d'un article de « Le Monde.fr » (31/10/07)

Pour la première fois depuis les affrontements qui les ont opposés à la junte militaire à la fin du mois de septembre, entre cent et deux cents bonzes ont défilé pacifiquement pendant plus d'une heure dans la ville de Pakko, au centre de la Birmanie, mercredi 31 novembre, selon des témoins.

Idée directrice de notre approche

Une phrase contient souvent de l'information peu importante

Question : peut-on compresser un texte en supprimant l'information moins importante des phrases ?

Exemple, extrait d'un article de « Le Monde.fr » (31/10/07)

Pour la première fois depuis les affrontements qui les ont opposés à la junte militaire à la fin du mois de septembre, entre cent et deux cents bonzes ont défilé pacifiquement pendant plus d'une heure dans la ville de Pakko, au centre de la Birmanie, mercredi 31 novembre, selon des témoins.

Idée directrice de notre approche

Une phrase contient souvent de l'information peu importante

Question : peut-on compresser un texte en supprimant l'information moins importante des phrases ?

Exemple, extrait d'un article de « Le Monde.fr » (31/10/07)

Pour la première fois depuis les affrontements qui les ont opposés à la junte militaire à la fin du mois de septembre, entre cent et deux cents bonzes ont défilé pacifiquement pendant plus d'une heure dans la ville de Pakko, au centre de la Birmanie, mercredi 31 novembre, selon des témoins.

Idée directrice de notre approche

Une phrase contient souvent de l'information peu importante

Question : peut-on compresser un texte en supprimant l'information moins importante des phrases ?

Exemple, extrait d'un article de « Le Monde.fr » (31/10/07)

Pour la première fois depuis les affrontements qui les ont opposés à la junte militaire à la fin du mois de septembre, entre cent et deux cents bonzes ont défilé *pacifiquement pendant plus d'une heure* dans la ville de Pakko, *au centre de la Birmanie*, mercredi 31 novembre, selon des témoins.

Idée directrice de notre approche

Une phrase contient souvent de l'information peu importante

Question : peut-on compresser un texte en supprimant l'information moins importante des phrases ?

Exemple, extrait d'un article de « Le Monde.fr » (31/10/07)

Pour la première fois depuis les affrontements qui les ont opposés à la junte militaire à la fin du mois de septembre, entre cent et deux cents bonzes ont défilé *pacifiquement pendant plus d'une heure* dans la ville de Pakko, *au centre de la Birmanie*, mercredi 31 novembre, *selon des témoins*.

Idée directrice de notre approche

Une phrase contient souvent de l'information peu importante

Question : peut-on compresser un texte en supprimant l'information moins importante des phrases ?

Exemple, extrait d'un article de « Le Monde.fr » (31/10/07)

Pour la première fois depuis les affrontements qui les ont opposés à la junte militaire à la fin du mois de septembre, entre cent et deux cents bonzes ont défilé *pacifiquement pendant plus d'une heure* dans la ville de Pakko, *au centre de la Birmanie*, mercredi 31 novembre, *selon des témoins*.

$\Rightarrow$ Entre cent et deux cents bonzes ont défilé dans la ville de Pakko, mercredi 31 novembre.

Hypothèse de recherche

La fonction syntaxique des constituants des phrases est un facteur conséquent dans l'évaluation de l'importance de ces constituants

Hypothèse de recherche

La fonction syntaxique des constituants des phrases est un facteur conséquent dans l'évaluation de l'importance de ces constituants

Exemple, extrait d'un article de « Le Monde.fr » (31/10/07)

Hypothèse de recherche

La fonction syntaxique des constituants des phrases est un facteur conséquent dans l'évaluation de l'importance de ces constituants

Exemple, extrait d'un article de « Le Monde.fr » (31/10/07)

Pour la première fois depuis les affrontements qui les ont opposés à la junte militaire à la fin du mois de septembre, entre cent et deux cents bonzes ont défilé pacifiquement pendant plus d'une heure dans la ville de Pakko, au centre de la Birmanie, mercredi 31 novembre, selon des témoins.

- sujet
- verbe
- complément d'objet (direct, indirect, second)
- complément circonstanciel (lieu, temps, moyen, manière...)

Plan de la présentation

- le résumé automatique de texte
- notre modèle de compression syntaxique
- COLIN : notre compresseur automatique ou interactif
- l'évaluation de notre approche
- la conclusion et les perspectives

Plan de la présentation

- le résumé automatique de texte
- notre modèle de compression syntaxique
- COLIN : notre compresseur automatique ou interactif
- l'évaluation de notre approche
- la conclusion et les perspectives

Le résumé automatique : les méthodes traditionnelles

Critère d'importance principal : dans un ensemble de documents, un mot important pour un document est un mot fréquent dans ce dernier mais pas dans les autres[1].

⇒ les mots seuls ne peuvent former un résumé cohérent
⇒ conserver les phrases qui possèdent le plus de mots importants

[1] métrique du quotien $tf \times idf$ de [Salton & Yang, 1973].

Le résumé automatique : les méthodes traditionnelles

Critère d'importance principal : dans un ensemble de documents, un mot important pour un document est un mot fréquent dans ce dernier mais pas dans les autres[1].

⇒ les mots seuls ne peuvent former un résumé cohérent
⇒ conserver les phrases qui possèdent le plus de mots importants

Caractéristiques

- ✓ méthode très abordable (informations de surface)
- ✓ conservation de la cohérence des phrases
- ✗ perte partielle de la cohérence du texte
- ✗ les phrases ne sont pas résumées

[1] métrique du quotien $tf \times idf$ de [Salton & Yang, 1973].

Le résumé automatique : les méthodes traditionnelles

Critère d'importance principal : dans un ensemble de documents, un mot important pour un document est un mot fréquent dans ce dernier mais pas dans les autres[1].

⇒ les mots seuls ne peuvent former un résumé cohérent
⇒ conserver les phrases qui possèdent le plus de mots importants

Caractéristiques

- ✓ méthode très abordable (informations de surface)
- ✓ conservation de la cohérence des phrases
- ✗ perte partielle de la cohérence du texte
- ✗ les phrases ne sont pas résumées

[1] métrique du quotien $tf \times idf$ de [Salton & Yang, 1973].

Exploitation de la structure du texte

Structure rhétorique [Mann & Thompson, 1987]

- relations entre éléments du discours
- exemples : arrière-plan, élaboration, contraste, reformulation
- relation entre noyau (important) et satellite (moins important)

Exploitation de la structure du texte

Structure rhétorique [Mann & Thompson, 1987]

- relations entre éléments du discours
- exemples : arrière-plan, élaboration, contraste, reformulation
- relation entre noyau (important) et satellite (moins important)

Dans le résumé automatique : [Marcu, 1998], [Sukvaree *et al.*, 2007]

- localiser puis supprimer les satellites

✓ cohérence du texte théoriquement conservée

✓ réduction possible des phrases

✗ structure rhétorique très difficilement extractible

Exploitation de la structure des phrases

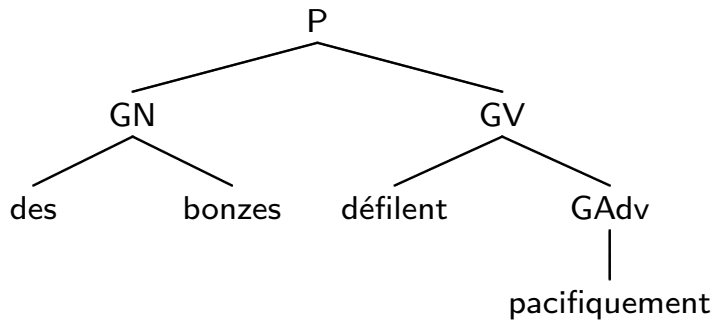
Structure syntaxique des phrases

- relations entre mots, ou mots et groupes de mots (constituants)

Exploitation de la structure des phrases

Structure syntaxique des phrases

- relations entre mots, ou mots et groupes de mots (constituants)
- exemple, syntaxe en constituants :



Exploitation de la structure des phrases

Structure syntaxique des phrases

Dans le résumé automatique

- ✓ conservation possible de la cohérence du texte
- ✓ réduction possible des phrases

Exploitation de la structure des phrases

Structure syntaxique des phrases

Dans le résumé automatique

- ✓ conservation possible de la cohérence du texte
- ✓ réduction possible des phrases

Deux principales approches statistiques :

- [Knight & Marcu, 2002]
- [Hovy *et al.*, 2005]

[Knight & Marcu, 2002]

2 exploitations différentes de la structure syntaxique

Un modèle de canal bruité

- hypothèse : les phrases du texte original furent autrefois courtes et on y a ajouté du bruit, de l'information moins importante
- objectif : retrouver le bruit et l'éliminer
- méthode : calcul de probabilités d'élagages (GHCP[2])

Un modèle fondé sur la décision

- hypothèse : les phrases du résumé sont issues de transformations sur celles du texte original
- objectif : appliquer des opérations similaires pour réaliser un résumé
- méthode : calcul de probabilités de transformations (GHCP)

[2] Grammaire Hors-Contexte Probabiliste

[Knight & Marcu, 2002]

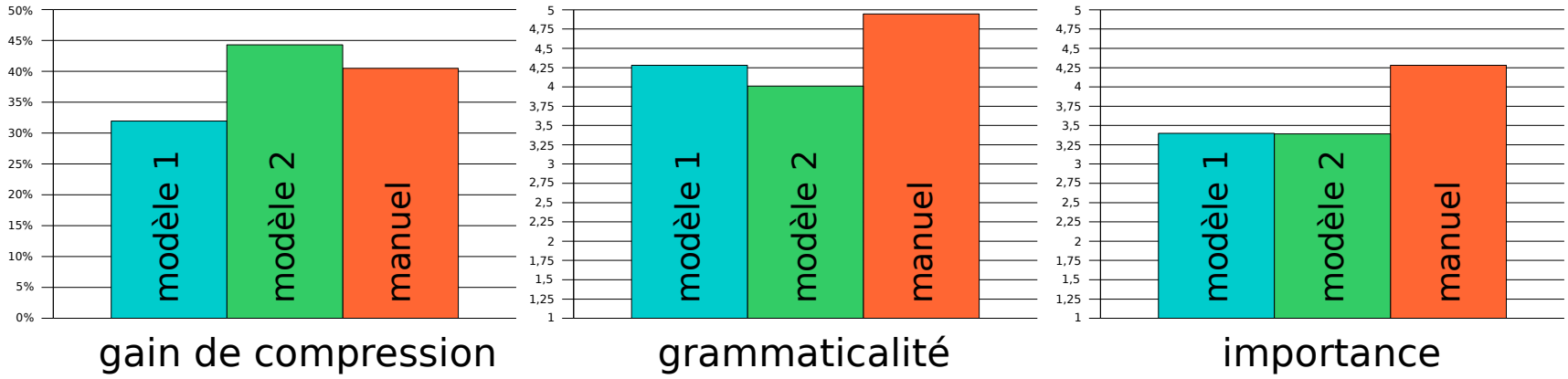
Caractéristiques des 2 modèles

- contraintes sur les performances
  - ✗ dépendance au corpus d'apprentissage (vocabulaire, cas syntaxiques)
  - ✗ dépendance au modèle probabiliste (paramètres limités et arbitraires)

[Knight & Marcu, 2002]

Caractéristiques des 2 modèles

- contraintes sur les performances
  - ✗ dépendance au corpus d'apprentissage (vocabulaire, cas syntaxiques)
  - ✗ dépendance au modèle probabiliste (paramètres limités et arbitraires)
- évaluation



[Hovy *et al.*, 2005]

Un modèle basé sur les dépendances et les arbres syntagmatiques

- hypothèse : les phrases peuvent se décomposer en relations de dépendances importantes ou pas
- objectif : localiser les relations peu importantes et retirer des phrases les parties impliquées
- méthode : importance des relations basée sur un rapport de vraisemblance, et grammaticalité sur une GHCP

[Hovy *et al.*, 2005]

Caractéristiques

- ✓ granularité d'importance minimale
  ⇒ relations de dépendances vs. arbres pour [Knight & Marcu, 2002]
- contraintes sur les performances
  - ✗ dépendance aux corpus d'apprentissage (relations importantes, cas syntaxiques)
  - ✗ dépendance aux modèles probabilistes et statistiques

[Hovy *et al.*, 2005]

Caractéristiques

- ✓ granularité d'importance minimale
  ⇒ relations de dépendances vs. arbres pour [Knight & Marcu, 2002]
- contraintes sur les performances
  - ✗ dépendance aux corpus d'apprentissage (relations importantes, cas syntaxiques)
  - ✗ dépendance aux modèles probabilistes et statistiques
- évaluation[3]
  - ✗ rappel faible
  - grammaticalité non mesurée

[3] méthode d'extraction et compression de phrases, tâche de résumé à taille fixe (100 mots), corpus de DUC2003.

Bilan et objectifs

modèle	*importance*	*grammaticalité*	*dépend de*
K.M. canal bruité	probabilités d'élagage et d'adjacence	probabilité de génération	corpus/modèle d'apprentissage
K.M. décision	probabilité de transformation		
H. *et al.*	fréquence des relations de dépendances		corpus/modèle d'apprentissage, modèle statistique

Bilan et objectifs

modèle	*importance*	*grammaticalité*	*dépend de*
K.M. canal bruité	probabilités d'élagage et d'adjacence	probabilité de génération	corpus/modèle d'apprentissage
K.M. décision	probabilité de transformation		
H. *et al.*	fréquence des relations de dépendances		corpus/modèle d'apprentissage, modèle statistique

$\Rightarrow$ **Notre objectif** : indépendance $\begin{cases} \text{à tout corpus d'apprentissage} \\ \text{à un modèle d'apprentissage} \end{cases}$

Bilan et objectifs

modèle	*importance*	*grammaticalité*	*dépend de*
K.M. canal bruité	probabilités d'élagage et d'adjacence	probabilité de génération	corpus/modèle d'apprentissage
K.M. décision	probabilité de transformation		
H. *et al.*	fréquence des relations de dépendances		corpus/modèle d'apprentissage, modèle statistique
notre objectif	propriétés syntaxiques		théorie linguistique

$\Rightarrow$ **Notre objectif** : indépendance $\begin{cases} \text{à tout corpus d'apprentissage} \\ \text{à un modèle d'apprentissage} \end{cases}$

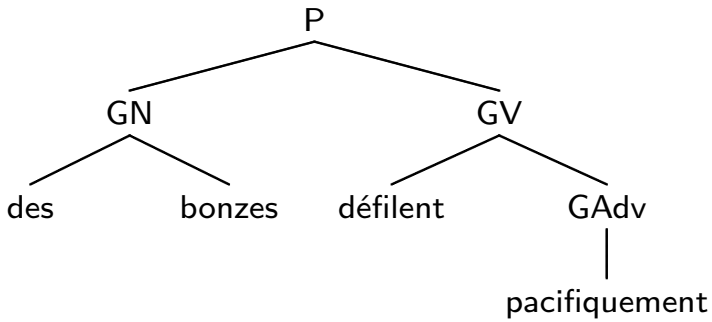
Plan de la présentation

- le résumé automatique de texte
- notre modèle de compression syntaxique
- COLIN : notre compresseur automatique ou interactif
- l'évaluation de notre approche
- la conclusion et les perspectives

Choix du modèle syntaxique

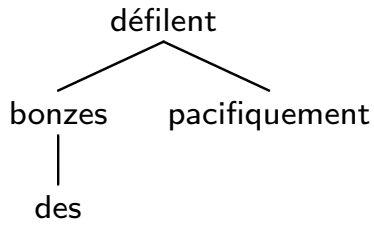
Deux principaux modèles de représentation

grammaire de constituants



- relation mot — groupe de mots
- [Chomsky, 1957]

grammaire de dépendance

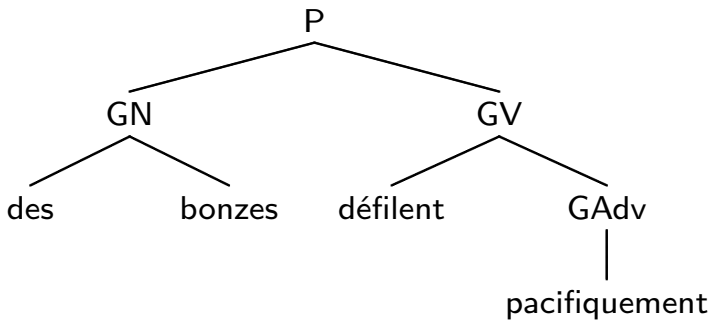


- relation mot — mot
- [Tesnière, 1934]

Choix du modèle syntaxique

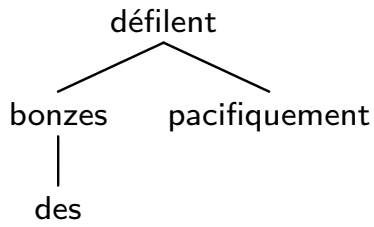
Deux principaux modèles de représentation

grammaire de constituants



- relation mot — groupe de mots
- [Chomsky, 1957]

grammaire de dépendance



- relation mot — mot
- [Tesnière, 1934]

La théorie du gouvernement et du liage [Chomsky, 1981]

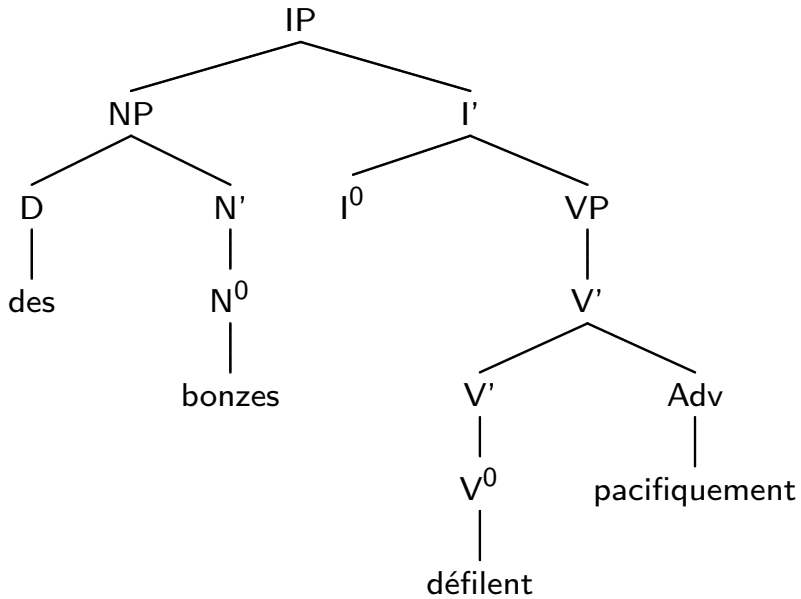
Système de règles structurelles génératives X-barre

$$
\begin{aligned}
XP &\rightarrow \text{spécifieur } X' \\
X' &\rightarrow X^0 \text{ compléments} \\
X' &\rightarrow X' \text{ conjonction } X' \\
X' &\rightarrow \text{adjoint } X'
\end{aligned}
$$

La théorie du gouvernement et du liage [Chomsky, 1981]

Système de règles structurelles génératives X-barre

$$
\begin{aligned}
XP &\rightarrow \text{spécifieur } X' \\
X' &\rightarrow X^0 \text{ compléments} \\
X' &\rightarrow X' \text{ conjonction } X' \\
X' &\rightarrow \text{adjoint } X'
\end{aligned}
$$



La théorie du gouvernement et du liage [Chomsky, 1981]

Tout constituant est gouverné par une tête

- un gouverné est moins important que sa tête
- un gouverné dispose d'une fonction selon son rôle et sa position dans la phrase
  $\Rightarrow$ fonction de spécifieur, complément ou adjoint

La théorie du gouvernement et du liage [Chomsky, 1981]

Tout constituant est gouverné par une tête

- un gouverné est moins important que sa tête
- un gouverné dispose d'une fonction selon son rôle et sa position dans la phrase
  $\Rightarrow$ fonction de spécifieur, complément ou adjoint

Équivalences avec la grammaire classique française

- spécifieur : déterminant, adverbe, sujet
- complément : COD, COI, complément du nom, groupe verbal
- adjoint : adjectif, complément circonstanciel

Organisation de la partie théorique

Organisation de la partie théorique

1. théorie linguistique de support
2. notre grammaire : conservation de la cohérence syntaxique

Notre grammaire basée sur X-barre

Adaptation de X-barre orientée vers l'importance syntaxique

Propriété d'effaçabilité

Caractéristiques inadaptées dans X-barre

- ✗ spécifieur : effaçabilité variable
  ⇒ déterminant et sujet toujours obligatoires
  ⇒ adverbes facultatifs
- ✗ complément : effaçabilité variable
  ⇒ groupe verbal toujours obligatoire
  ⇒ autres compléments importants selon leur tête (sous-catégorisation)
- ✗ adjoint verbal : portée sémantique variable
  ⇒ compléments circonstanciels non sous-catégorisés
- ✗ têtes peu considérées : pronom, adverbe

Notre solution

définition de 2 classes de fonctions selon l'effaçabilité :
les compléments et les modifieurs

Notre solution

définition de 2 classes de fonctions selon l'effaçabilité :
les compléments et les modifieurs

Les compléments

- éléments sous-catégorisés par leur tête
  $\Rightarrow$ compléments obligatoires ou non

Notre solution

définition de 2 classes de fonctions selon l'effaçabilité :
les compléments et les modifieurs

Les compléments

- éléments sous-catégorisés par leur tête
  $\Rightarrow$ compléments obligatoires ou non

Les modifieurs

- éléments effaçables quelque soit leur tête

Notre solution

définition de 2 classes de fonctions selon l'effaçabilité :
les compléments et les modifieurs

Les compléments

- éléments sous-catégorisés par leur tête
  ⇒ compléments obligatoires ou non

Les modifieurs

- éléments effaçables quelque soit leur tête

⇒ répartition des cas syntaxiques de X-barre dans nos classes[4]

[4] adaptation pour le français.

Les compléments (en français)

Nous incluons dans cette classe :

- le déterminant
  ⇒ complément obligatoire du nom commun
- le groupe sujet
  ⇒ complément obligatoire de la proposition
- le groupe verbal
  ⇒ complément obligatoire de la proposition
- le complément de l'adverbe
  ⇒ effaçabilité dépend de sa tête
  ⇒ ex, non effaçable : *Les bonzes agissent indépendamment de la volonté de la junte militaire.*
- le complément du pronom
  ⇒ effaçabilité dépend de sa tête
  ⇒ ex, effaçable : *Lesquels des bonzes braveront la peur de la junte militaire ?*

Les modifieurs (en français)

Nous incluons dans cette classe :

- le groupe adverbial

  $\Rightarrow$ cas de spécifieurs dans X-barre

- les compléments circonstanciels non sous-catégorisés par le verbe

  $\Rightarrow$ cas des adjoints du verbe dans X-barre

  $\Rightarrow$ nous le plaçons en gouverné par la proposition

  $\Rightarrow$ ex : *Les bonzes ont défilé pendant plus d'une heure*.

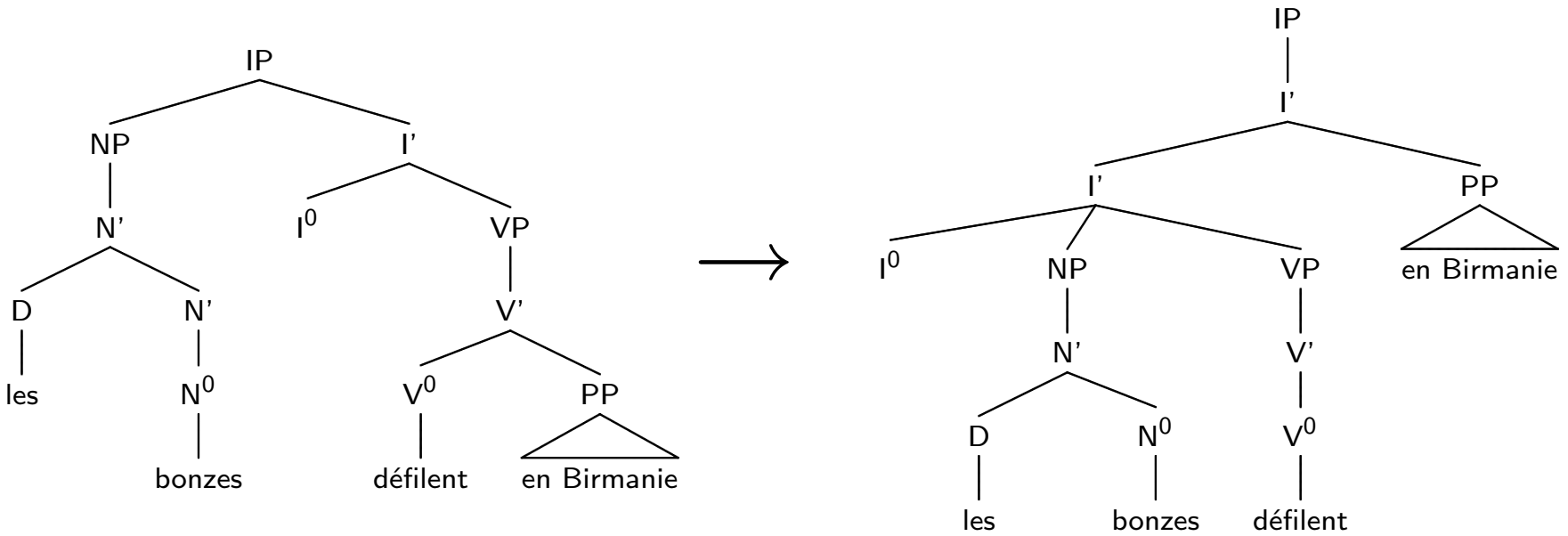
Modification du système de règles structurelles

$$
\begin{aligned}
XP &\rightarrow \text{spécifieur } X' \\
X' &\rightarrow X^0 \text{ compléments} \\
X' &\rightarrow X' \text{ conjonction } X' \\
X' &\rightarrow \text{adjoint } X'
\end{aligned}
$$

$$
\begin{aligned}
XP &\rightarrow X' \\
X' &\rightarrow \text{compléments } X^0 \text{ compléments} \\
X' &\rightarrow X' \text{ conjonction } X' \\
X' &\rightarrow \text{modifieur } X' \\
X' &\rightarrow X' \text{ modifieur}
\end{aligned}
$$

Exemple de modfication d'arbre syntaxique

Les bonzes défilent en Birmanie



Effaçabilité, récapitulatif

tête	*complément*	*modifieur*
nom	✓/✗	✓
pronom	✓/✗	✓
adjectif	✓/✗	✓
verbe	✓/✗	✓
adverbe	✓/✗	✓
préposition	✗	✓
proposition	✗	✓

Effaçabilité, récapitulatif

tête	*complément*	*modifieur*
nom	✓/✗	✓
pronom	✓/✗	✓
adjectif	✓/✗	✓
verbe	✓/✗	✓
adverbe	✓/✗	✓
préposition	✗	✓
proposition	✗	✓

Quand effacer le complément ?

$\Rightarrow$ selon les traits de sous-catégorisation de la tête

$\Rightarrow$ exploitation d'une ressource lexicale

Organisation de la partie théorique

Organisation de la partie théorique

1. théorie linguistique de support
2. notre grammaire : conservation de la cohérence syntaxique
3. importance syntaxique : une ressource lexicale adéquate

Ressource lexicale de sous-catégorisaton

Propriétés requises

- ressource numérique exploitable automatiquement
- pour chaque tête lexicale, informations sur :
  - arité du prédicat
  - compléments obligatoires/facultatifs
  - propriétés morphologiques des compléments
    $\Rightarrow$ prépositions
    $\Rightarrow$ catégorie grammaticale
    $\Rightarrow$ ...

Le Le*fff*, ressource adéquate

Caractéristiques du Le*fff*

- ressource lexicale des formes fléchies du français [Sagot *et al.*, 2005]
- basée sur le lexique-grammaire (Maurice Gross)

Le Lefff, ressource adéquate

Caractéristiques du Lefff

- ressource lexicale des formes fléchies du français [Sagot *et al.*, 2005]
- basée sur le lexique-grammaire (Maurice Gross)

Entrées

version	*nombre de formes*	*nombre de lemmes*
2.1	404 634	105 595

Exemples de champs

#	*forme fléchie*	*cat.*	*traits de sous-catégorisation*
1	coupe	v	<suj:(sn),obj:(sn)>
2	coupe	v	<suj:(sn|sinf|scompl),obja:à-sn|à-scompl|à-sinf>

Le Le*fff*, ressource adéquate

Exploitation du Le*fff*

Exemples de champs

#	*forme fléchie*	*cat.*	*traits de sous-catégorisation*
1	coupe	v	<suj:(sn),obj:(sn)>
2	coupe	v	<suj:(sn|sinf|scompl),obja:à-sn|à-scompl|à-sinf>

Marie coupe un gâteau.

Le Lefff, ressource adéquate

Exploitation du Lefff

Exemples de champs

#	*forme fléchie*	*cat.*	*traits de sous-catégorisation*
1	coupe	v	<suj:(sn),obj:(sn)>
2	coupe	v	<suj:(sn|sinf|scompl),obja:à-sn|à-scompl|à-sinf>

Marie coupe un gâteau. ⇒ entrée 1, complément effaçable

Le Lefff, ressource adéquate

Exploitation du Lefff

Exemples de champs

#	*forme fléchie*	*cat.*	*traits de sous-catégorisation*
1	coupe	v	<suj:(sn),obj:(sn)>
2	coupe	v	<suj:(sn|sinf|scompl),obja:à-sn|à-scompl|à-sinf>

Marie coupe un gâteau. ⇒ entrée 1, complément effaçable
Marie coupe à l'école.

Le Lefff, ressource adéquate

Exploitation du Lefff

Exemples de champs

#	*forme fléchie*	*cat.*	*traits de sous-catégorisation*
1	coupe	v	<suj:(sn),obj:(sn)>
2	coupe	v	<suj:(sn|sinf|scompl),obja:à-sn|à-scompl|à-sinf>

Marie coupe un gâteau. ⇒ entrée 1, complément effaçable
Marie coupe à l'école. ⇒ entrée 2, complément obligatoire

Le Lefff, limitations actuelles pour notre approche

- sous-catégorisation précise des verbes partielle
- compléments obligatoires des adjectifs
  $\Rightarrow$ *Marie est encline à partir*
- pas de sous-catégorisation des adverbes
  $\Rightarrow$ *Marie agit conformément à la loi*

Le Lefff, limitations actuelles pour notre approche

- sous-catégorisation précise des verbes partielle
- compléments obligatoires des adjectifs
  ⇒ *Marie est encline à partir*
- pas de sous-catégorisation des adverbes
  ⇒ *Marie agit conformément à la loi*

Conclusion

✓ format et type des informations adéquats à notre approche

✗ ressource peu exploitable dans l'état courant

- stratégie de prudence pour les compléments
  ⇒ par défaut, conserver les compléments
  ⇒ selon le taux de compression requis, supprimer les compléments

Organisation de la partie théorique

Organisation de la partie théorique

1. théorie linguistique de support
2. notre grammaire : conservation de la cohérence syntaxique
3. importance syntaxique : une ressource lexicale adéquate
4. importance sémantique : des indices linguistiques

Les Fonctions Lexicales

Fonction Lexicale : relation de proximité sémantique et de co-textualité entre lemmes

Les Fonctions Lexicales

Fonction Lexicale : relation de proximité sémantique et de co-textualité entre lemmes

Exemple, importance d'un complément du nom

- *on peut maintenant traiter le cœur du problème.*
  ⇒ FL « nom du centre » entre *cœur* et *du problème*
  ⇒ le complément est important

Les Fonctions Lexicales

Fonction Lexicale : relation de proximité sémantique et de co-textualité entre lemmes

Exemple, importance d'un complément du nom

- *on peut maintenant traiter le cœur du problème.*
  ⇒ FL « nom du centre » entre *cœur* et *du problème*
  ⇒ le complément est important
- *on peut maintenant manger le cœur du poulet.*
  ⇒ pas de FL ici
  ⇒ le complément est moins important

Les Fonctions Lexicales

Fonction Lexicale : relation de proximité sémantique et de co-textualité entre lemmes

Exemple, importance d'un complément du nom

- *on peut maintenant traiter le cœur du problème.*
  ⇒ FL « nom du centre » entre *cœur* et *du problème*
  ⇒ le complément est important
- *on peut maintenant manger le cœur du poulet.*
  ⇒ pas de FL ici
  ⇒ le complément est moins important

Autres exemples :

Figuratif	→	*rideau de fumée*
Singulatif	→	*grain de riz / de sel / de folie*
Collectif	→	*flotte de navires / de bateaux*
Dérivé sémantique actanciel	→	*rempli, plein de mépris*

Ressource exploitable pour les FL

le DiCo

- basé sur le Dictionnaire Explicatif et Combinatoire [Mel'čuk *et al.*, 1995]
- ✗ 1075 lemmes[5] (acceptions)
  ⇒ petite partie du lexique français (300000 dans le TLFi)

[5] au 1er août 2007

Ressource exploitable pour les FL

le DiCo

- basé sur le Dictionnaire Explicatif et Combinatoire [Mel'čuk *et al.*, 1995]
- ✗ 1075 lemmes[5] (acceptions)
  ⇒ petite partie du lexique français (300000 dans le TLFi)

Conclusion : exploitation possible mais peu rentable à l'heure actuelle

[5] au 1er août 2007

Autres indices linguistiques

- phrasèmes complets
  ⇒ *Marie a bu la tasse*
- type du déterminant
  ⇒ article indéfini : *Un grand chien a mangé le chat de Marie*
  ⇒ article défini : *Le grand chien a mangé le chat de Marie* (pas le petit)
- éléments incidents
  ⇒ *entre cent et deux cents bonzes ont défilé, selon des témoins*
- modifieur du nom détaché
  ⇒ *la junte militaire étoffe son armée, frappée par des desertions*
- position du constituant dans la phrase
  ⇒ « [Placés] en tête de phrase, en position de thème, [...] [les] circonstants de phrase constituent le cadre qui organise la cohérence du texte » [Tomassone, 2001]
- négation (importante)
- interrogation
  ⇒ *il travaille à Paris*
  ⇒ *travaille-t-il à Paris* ?

Organisation de la partie théorique

Organisation de la partie théorique

1. théorie linguistique de support
2. notre grammaire : conservation de la cohérence syntaxique
3. importance syntaxique : une ressource lexicale adéquate
4. importance sémantique : des indices linguistiques
5. influence du genre textuel

Influence du genre textuel

« le lexique, la morphosyntaxe, la manière dont se posent les problèmes sémantiques de l'ambiguïté et de l'implicite, tout cela varie avec les genres » [Rastier, 2002]

Influence du genre textuel

« le lexique, la morphosyntaxe, la manière dont se posent les problèmes sémantiques de l'ambiguïté et de l'implicite, tout cela varie avec les genres » [Rastier, 2002]

Tests manuels réalisés sur différents genres :

- textes narratifs (romans, contes, . . .)
- corpus de dépêches journalistiques
- articles scientifiques en biologie

Influence du genre textuel

« le lexique, la morphosyntaxe, la manière dont se posent les problèmes sémantiques de l'ambiguïté et de l'implicite, tout cela varie avec les genres » [Rastier, 2002]

Tests manuels réalisés sur différents genres :

- textes narratifs (romans, contes, . . .)
- corpus de dépêches journalistiques
- articles scientifiques en biologie

Premiers résultats estimés :

- genre narratif adéquat à la compression
- genre journalistique plus résistant
- genre scientifique peu adéquat

Influence du genre textuel

« le lexique, la morphosyntaxe, la manière dont se posent les problèmes sémantiques de l'ambiguïté et de l'implicite, tout cela varie avec les genres » [Rastier, 2002]

Tests manuels réalisés sur différents genres :

- textes narratifs (romans, contes, . . .)
- corpus de dépêches journalistiques
- articles scientifiques en biologie

Premiers résultats estimés :

- genre narratif adéquat à la compression
- genre journalistique plus résistant
- genre scientifique peu adéquat

$\Rightarrow$ le genre a un impact réel sur l'importance des constituants

$\Rightarrow$ l'expérimentation nous aidera à mieux l'estimer

Organisation de la partie théorique

Organisation de la partie théorique

1. théorie linguistique de support
2. notre grammaire : conservation de la cohérence syntaxique
3. importance syntaxique : une ressource lexicale adéquate
4. importance sémantique : des indices linguistiques
5. influence du genre textuel
6. limites de l'approche automatique

Les limites de l'approche automatique

Constats

- ressources lexicales actuellement très partielles
- autres informations extractibles du document
  $\Rightarrow$ autres informations au sein de la phrase
  $\Rightarrow$ informations contextuelles
- autres informations hors document
  $\Rightarrow$ les préférences du récepteur du résumé

Les limites de l'approche automatique

Constats

- ressources lexicales actuellement très partielles
- autres informations extractibles du document
  $\Rightarrow$ autres informations au sein de la phrase
  $\Rightarrow$ informations contextuelles
- autres informations hors document
  $\Rightarrow$ les préférences du récepteur du résumé

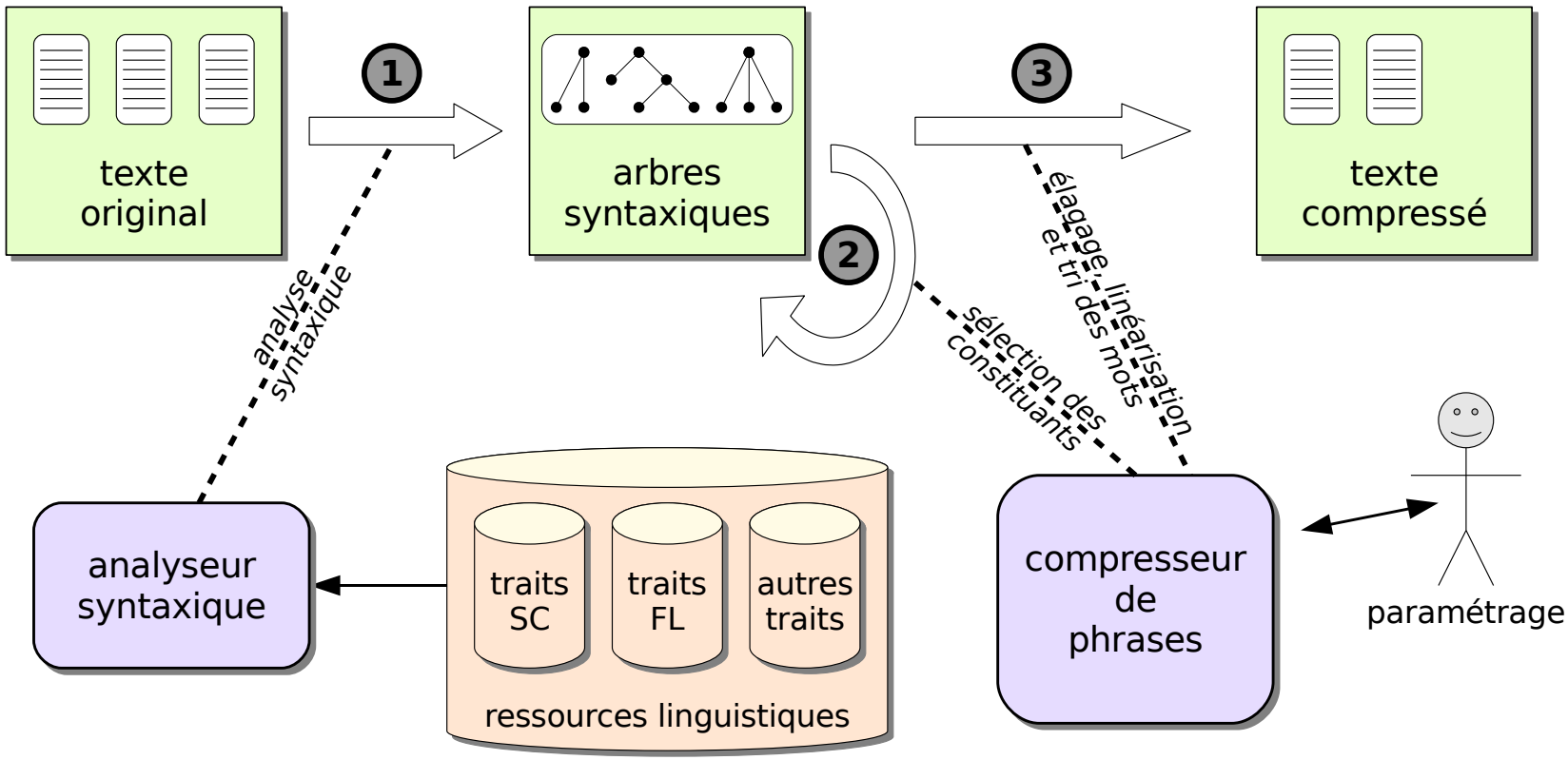
Exploration d'une piste parallèle

Compression interactive faisant intervenir les choix du récepteur

Plan de la présentation

- le résumé automatique de texte
- notre modèle de compression syntaxique
- COLIN : notre compresseur automatique ou interactif
- l'évaluation de notre approche
- la conclusion et les perspectives

Architecture de notre modèle de compression syntaxique



SYGFRAN, l'analyseur syntaxique

SYGMART[6] : décomposition puis transformation d'une chaîne textuelle en une structure arborescente

- modèle d'analyse calculatoire
- lexique généré exhaustif
- règles sous-contexte « conditions $\Rightarrow$ actions »
  $\Rightarrow$ schéma de reconnaissance structurel / schéma de transformation
- compléxité en $O(nlog_2(n))$, n de l'ordre du nombre de mots

[6] [Chauché, 1984]

SYGFRAN, l'analyseur syntaxique

SYGMART[6] : décomposition puis transformation d'une chaîne textuelle en une structure arborescente

- modèle d'analyse calculatoire
- lexique généré exhaustif
- règles sous-contexte « conditions ⇒ actions »
  ⇒ schéma de reconnaissance structurel / schéma de transformation
- compléxité en $O(nlog_2(n))$, n de l'ordre du nombre de mots

SYGFRAN : ensemble de règles pour SYGMART

- grammaire de constituants, très proche de la notre
- 20000 règles, 250 grammaires, 25000 lemmes
- estimation à 35 % d'analyses complètes (de phrases)
  ⇒ sinon, analyse partielle exploitable

[6] [Chauché, 1984]

La mise en œuvre : CoLin

CoLin

- notre outil de résumé automatique ou semi-automatique
- incluant une interface interactive
- basé sur l'analyseur Sygfran
- composé d'un ensemble de règles sous-contexte Sygmart

La mise en œuvre : CoLin

CoLin

- notre outil de résumé automatique ou semi-automatique
- incluant une interface interactive
- basé sur l'analyseur Sygfran
- composé d'un ensemble de règles sous-contexte Sygmart
  1. d'adaptation de la grammaire Sygfran
  2. de résolution de liens anaphoriques
  3. d'étiquetage des nœuds de constituants candidats à l'effacement
  4. de linéarisation des structures arborescentes

  $\Rightarrow$ production d'un texte étiqueté pour l'interface

Interface de CoLin

Caractéristiques

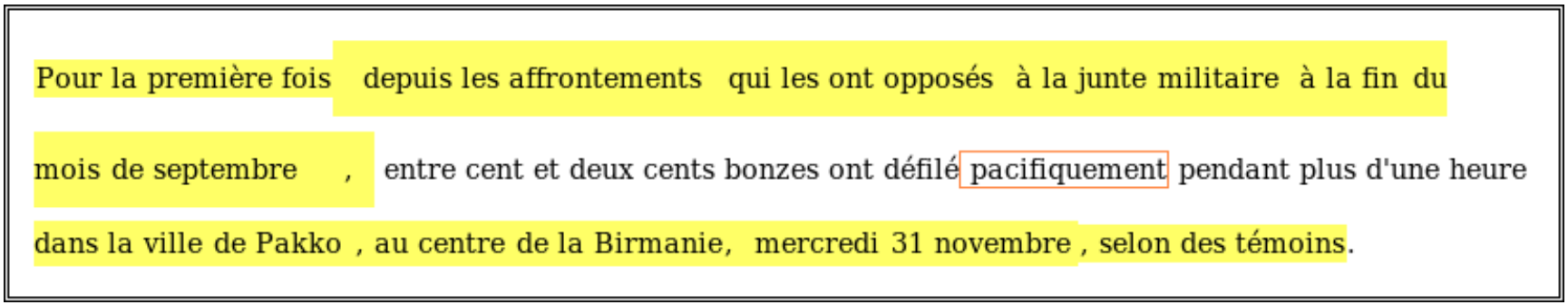
- interaction sur chaque constituant candidat
- code de couleur selon l'importance estimée

Interface de CoLin

Caractéristiques

- interaction sur chaque constituant candidat
- code de couleur selon l'importance estimée

Captures

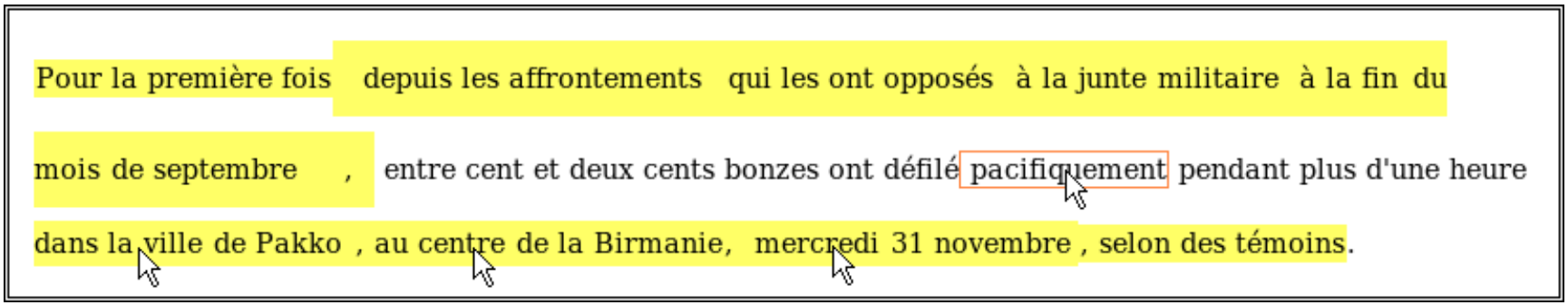


Interface de CoLin

Caractéristiques

- interaction sur chaque constituant candidat
- code de couleur selon l'importance estimée

Captures

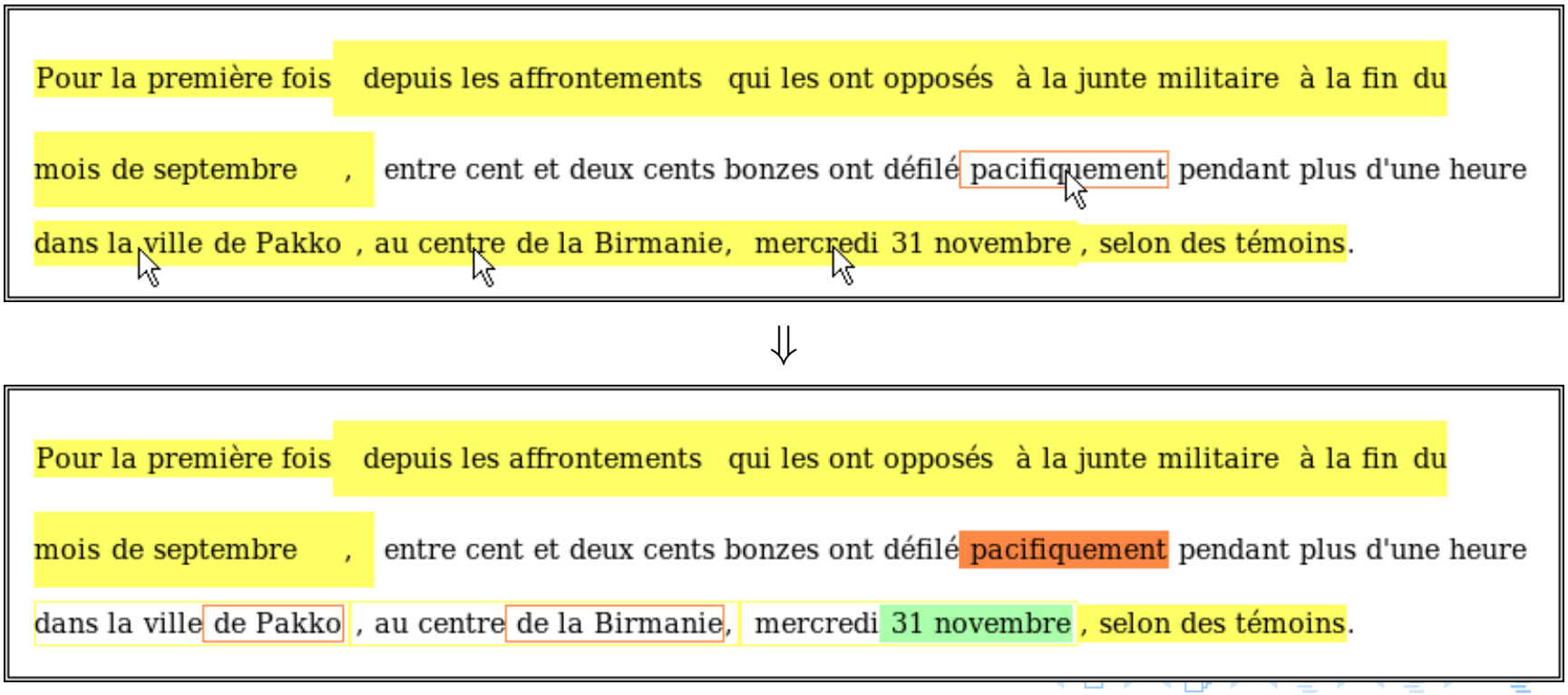


Interface de CoLin

Caractéristiques

- interaction sur chaque constituant candidat
- code de couleur selon l'importance estimée

Captures



Plan de la présentation

- le résumé automatique de texte
- notre modèle de compression syntaxique
- COLIN : notre compresseur automatique ou interactif
- l'évaluation de notre approche
- la conclusion et les perspectives

Protocole d'évaluation

Évaluation manuelle intrinsèque

inspirée de celle de [Knight & Marcu, 2002]

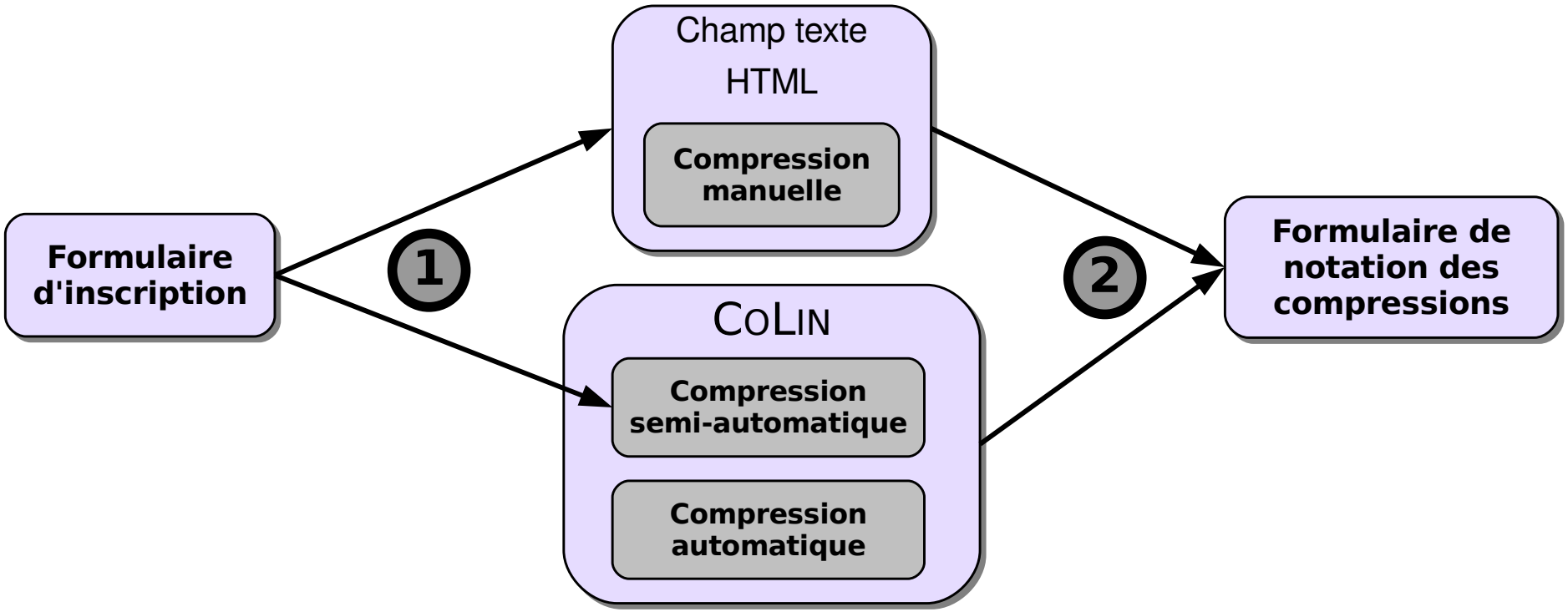
Protocole d'évaluation

Évaluation manuelle intrinsèque
inspirée de celle de [Knight & Marcu, 2002]

Objectifs

1. éviter la notation par comparaison avec résumé humain
   ⇒ éviter des scores de n-grammes
   ⇒ réduire la subjectivité
2. réduire l'effort cognitif des évaluateurs
   ⇒ maximiser la participation
3. évaluer la qualité des compressions
4. évaluer le temps gagné par la version semi-automatique
5. évaluer la satisfaction d'interaction
6. évaluer l'influence du genre textuel

Déroulement de l'évaluation



Étape 1

- affectation de la tâche
- affectation de 5 documents

Étape 2

- décomposition des documents en paragraphes
- affectation de 5 documents à noter
- affectation de 3 compressions par paragraphe

Les données de l'évaluation

Le corpus

- 3 genres : journalistique, narratif, scientifique
- 5 textes par genre
- 400 mots par texte
- analyse syntaxique du corpus améliorée manuellement

Les données de l'évaluation

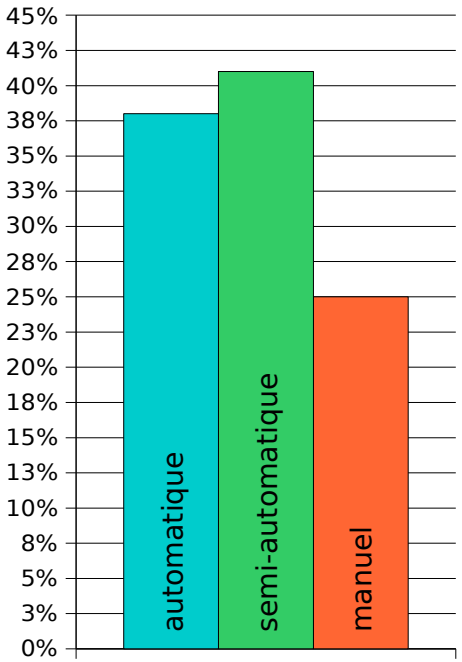
Le corpus

- 3 genres : journalistique, narratif, scientifique
- 5 textes par genre
- 400 mots par texte
- analyse syntaxique du corpus améliorée manuellement

Le panel et la participation

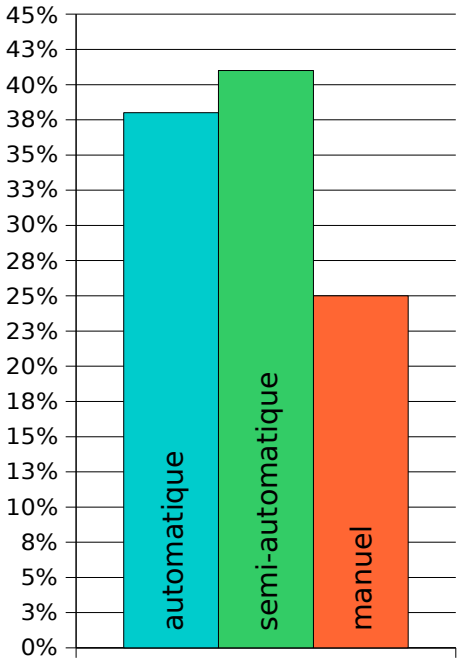
- 25 utilisateurs actifs
- 19 pour les compressions
- 18 pour les notations
- majorité de doctorants et docteurs, en informatique
- 3 à 4 compressions par document et par mode de compression
- 5 notes par paragraphe compressé

Compression, qualité, temps moyens (3 genres)

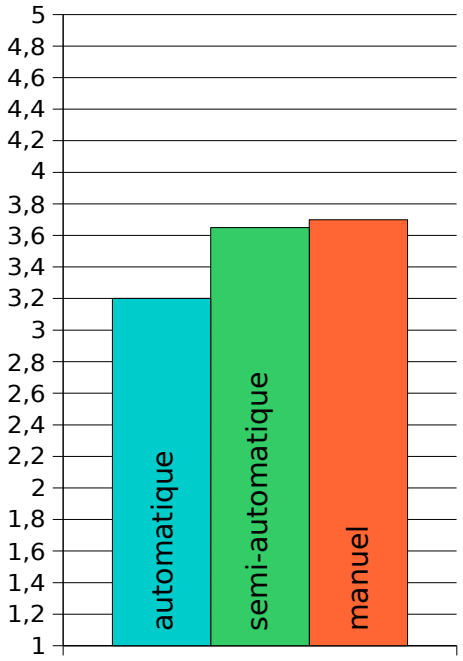


Gain de compression
(1-résumé/original)

Compression, qualité, temps moyens (3 genres)

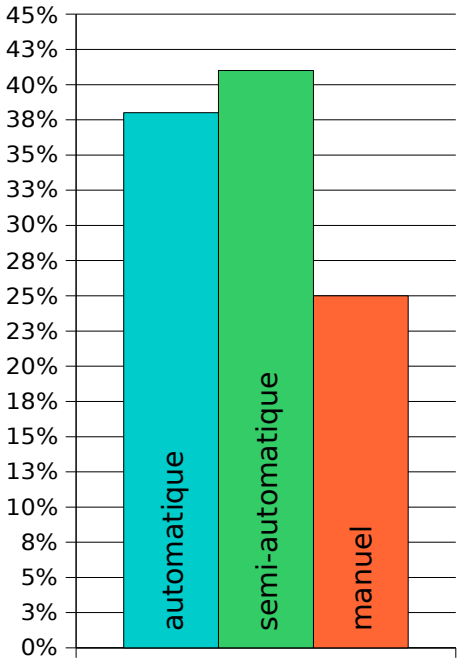


Gain de compression
(1-résumé/original)

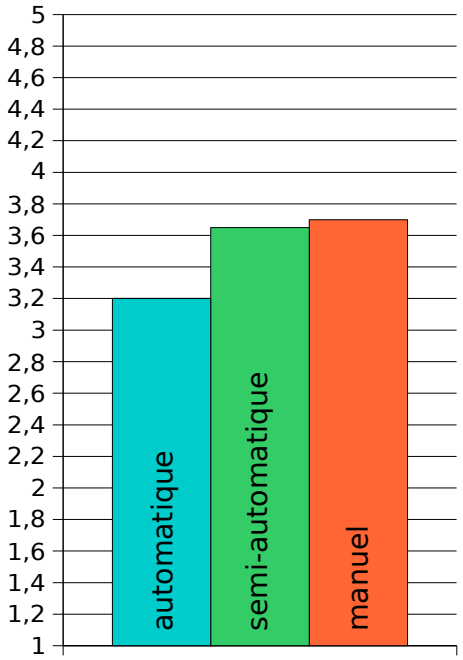


Qualité de compression
(1<note<5)

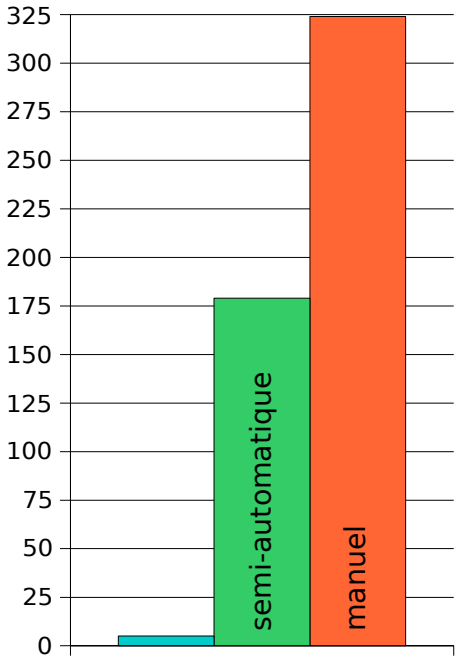
Compression, qualité, temps moyens (3 genres)



Gain de compression
(1-résumé/original)

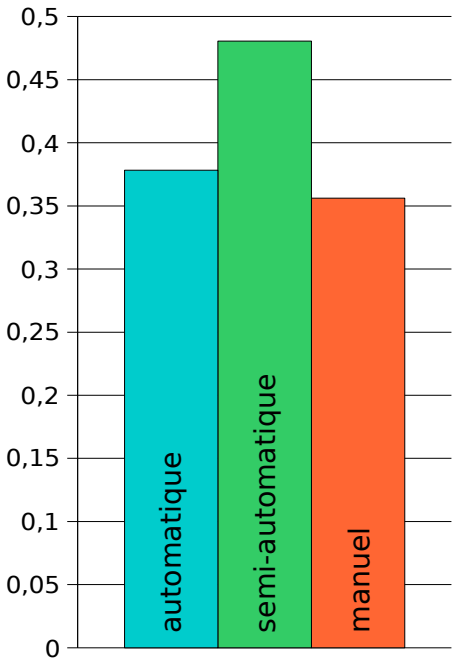


Qualité de compression
(1<note<5)

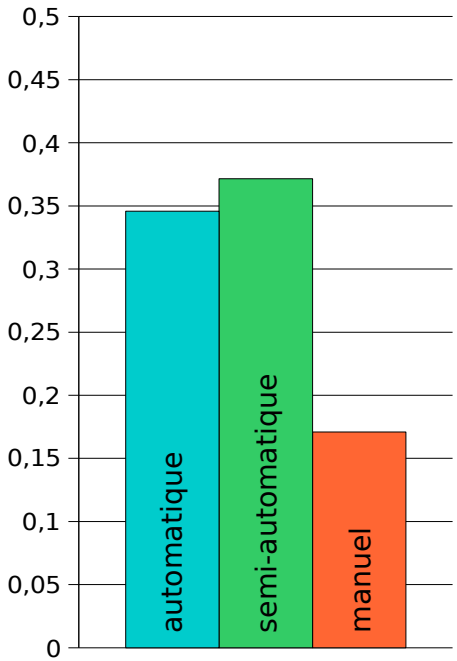


Temps de compression
(secondes, 5ème texte)

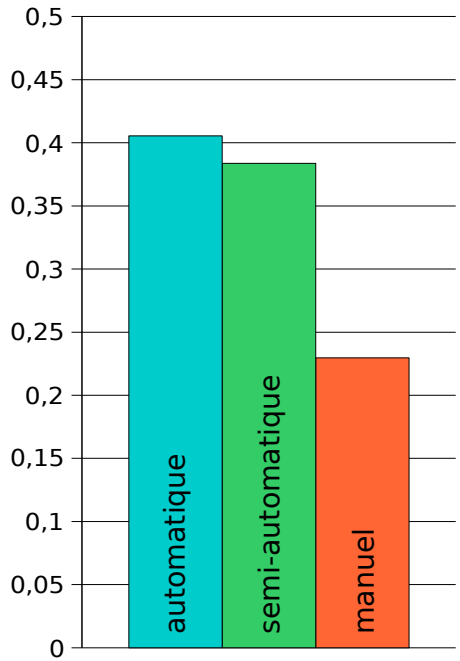
Gain de compression selon le genre



Corpus journalistique



Corpus narratif



Corpus scientifique

Bilan de l'évaluation

- ✓ résumés automatique et semi-automatique très satisfaisants
  $\Rightarrow$ malgré une analyse syntaxique imparfaite
- ✓ évaluateurs satisfaits de leur interaction (note $\approx 4/5$)
- ✓ genre journalistique propice à la compression
- ✓ genre scientifique autant propice à la compression que le narratif !
- ✓ qualité des résumés subjective
  $\Rightarrow$ justification du protocole manuel
- ✓ amélioration de SYGFRAN

Plan de la présentation

- le résumé automatique de texte
- notre modèle de compression syntaxique
- COLIN : notre compresseur automatique ou interactif
- l'évaluation de notre approche
- la conclusion et les perspectives

Bilan

Résumé par compression syntaxique basé sur des critères linguistiques

Bilan

Résumé par compression syntaxique basé sur des critères linguistiques

- compression moyenne $\approx$ 40 %
  $\Rightarrow$ applications particulières
  $\Rightarrow$ sous-tâche d'un résumé automatique
- nécessite une analyse syntaxique correcte
  $\Rightarrow$ analyse partielle exploitable en semi-automatique

Bilan

Résumé par compression syntaxique basé sur des critères linguistiques

- compression moyenne $\approx$ 40 %
  $\Rightarrow$ applications particulières
  $\Rightarrow$ sous-tâche d'un résumé automatique
- nécessite une analyse syntaxique correcte
  $\Rightarrow$ analyse partielle exploitable en semi-automatique

✓ granularité du constituant adaptée

✓ impact réel des propriétés syntaxiques sur l'importance des constituants

Contributions

Modèle théorique computationnel validé

- méthode de compression syntaxique basée sur des propriétés linguistiques théoriques et empiriques

Contributions

Modèle théorique computationnel validé

- méthode de compression syntaxique basée sur des propriétés linguistiques théoriques et empiriques

Outil de compression de texte performant

- mode automatique
- mode semi-automatique par interaction locale inédite

Contributions

Modèle théorique computationnel validé

- méthode de compression syntaxique basée sur des propriétés linguistiques théoriques et empiriques

Outil de compression de texte performant

- mode automatique
- mode semi-automatique par interaction locale inédite

Protocole d'évaluation de résumé par compression de phrases adéquat

- évaluation manuelle intrinsèque
- facilitation de la tâche des évaluateurs
- limitation de la subjectivité de l'évaluation

Perspectives

Amélioration sur la sous-catégorisation

- enrichissement et intégration des données du Le*fff*
  $\Rightarrow$ sous-catégorisation des adverbes et adjectifs

Perspectives

Amélioration sur la sous-catégorisation

- enrichissement et intégration des données du Le*fff*
  $\Rightarrow$ sous-catégorisation des adverbes et adjectifs

Apprentissage sur l'interaction

- exploiter l'interaction pour améliorer la sélection des constituants
  $\Rightarrow$ critères : FS, position dans la phrase, type de circonstant, nature de l'entité, propriétés des constituants parents ou frères...

Perspectives

Amélioration sur la sous-catégorisation

- enrichissement et intégration des données du Le*fff*
  ⇒ sous-catégorisation des adverbes et adjectifs

Apprentissage sur l'interaction

- exploiter l'interaction pour améliorer la sélection des constituants
  ⇒ critères : FS, position dans la phrase, type de circonstant, nature de l'entité, propriétés des constituants parents ou frères...

La compression comme sous-tâche

Exploitation conjointe avec une compression de la structure

- rhétorique
- thématique
- ...

Fin

Question récurrente

Fin

Question récurrente

Compression automatique ou semi-automatique de textes par élagage des constituants effaçables : une approche interactive et indépendante des corpus

Fin

Question récurrente

Compression automatique ou semi-automatique de textes par élagage des constituants effaçables : une approche interactive et indépendante des corpus

Compression de textes

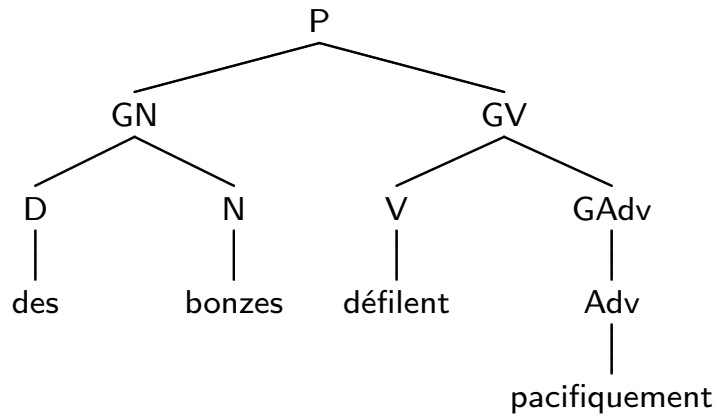
Fin

Merci de votre attention

[Knight & Marcu, 2002] : un modèle de canal bruité

1. Analyse syntaxique, arbres en constituants [Collins, 1997]

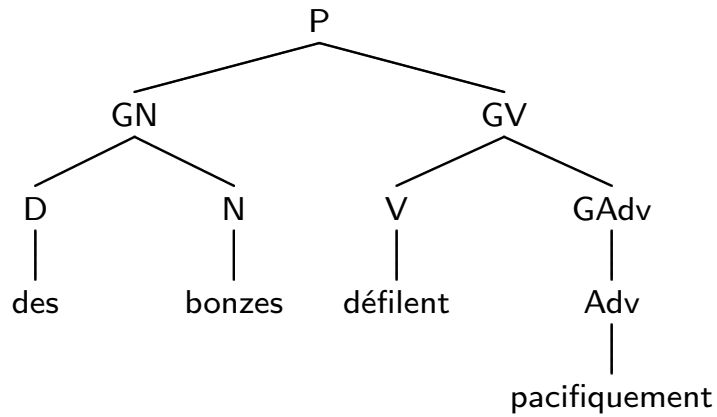
des bonzes défilent pacifiquement $\longrightarrow$



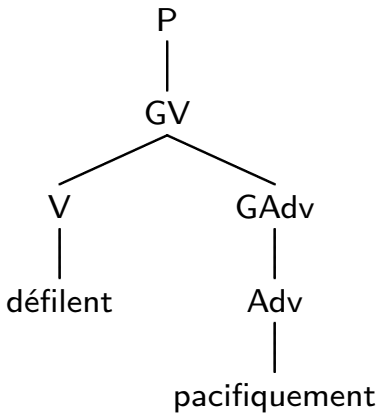
[Knight & Marcu, 2002] : un modèle de canal bruité

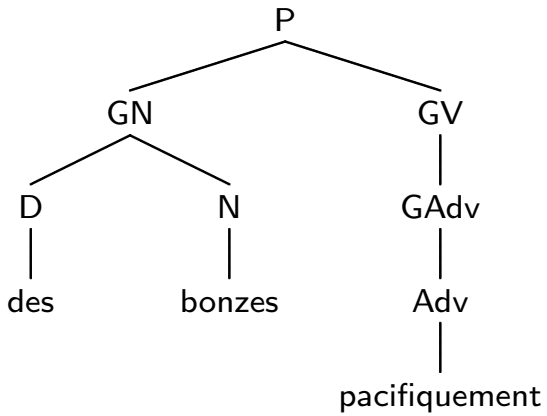
1. Analyse syntaxique, arbres en constituants [Collins, 1997]

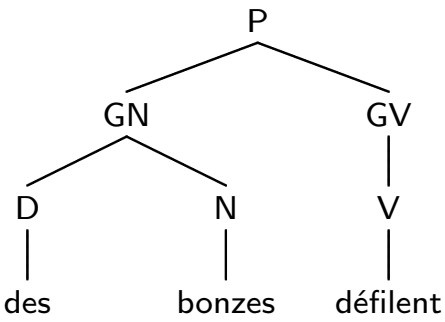
des bonzes défilent pacifiquement $\longrightarrow$



2. Génération de sous-arbres prétendants







[Knight & Marcu, 2002] : un modèle de canal bruité

3. par phrase, choix de l'arbre le plus probable

[Knight & Marcu, 2002] : un modèle de canal bruité

3. par phrase, choix de l'arbre le plus probable
   - entraînement : corpus (Ziff–Davis) de documents d'actualités associés à leur résumé

[Knight & Marcu, 2002] : un modèle de canal bruité

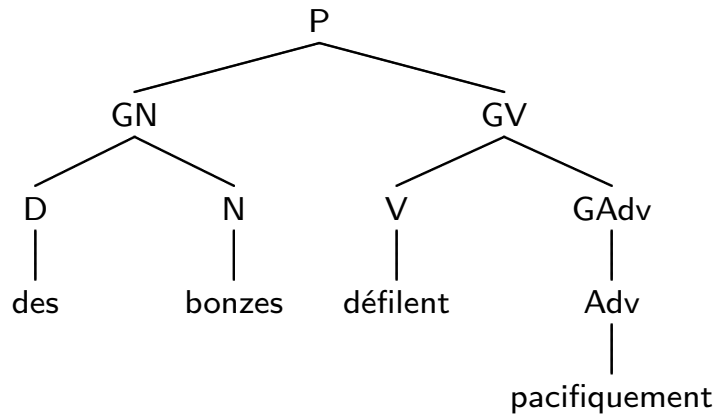
3. par phrase, choix de l'arbre le plus probable
   - entraînement : corpus (Ziff–Davis) de documents d'actualités associés à leur résumé
   - par nœud, score de Grammaires Hors-Contexte Probabilistes

   $\Rightarrow$ grammaticalité, ex : $P_{GHC}(P \rightarrow GN\ GV|P)$, $P_{GHC}(V \rightarrow \textit{défilent}|V)$

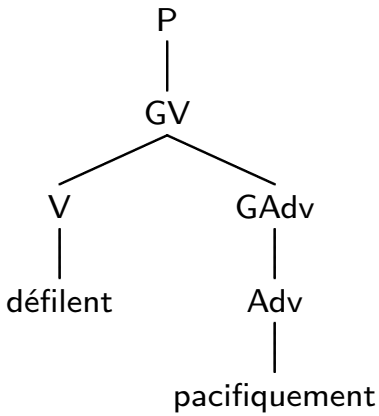
[Knight & Marcu, 2002] : un modèle de canal bruité

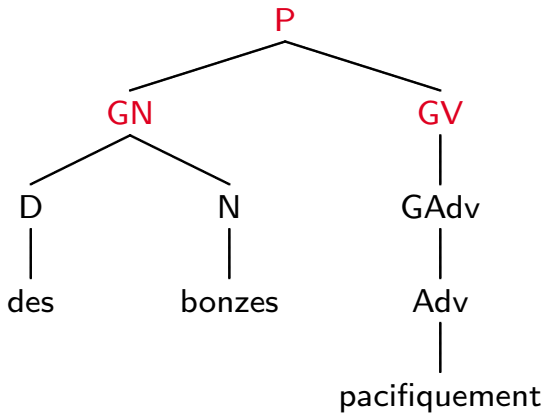
1. Analyse syntaxique, arbres en constituants [Collins, 1997]

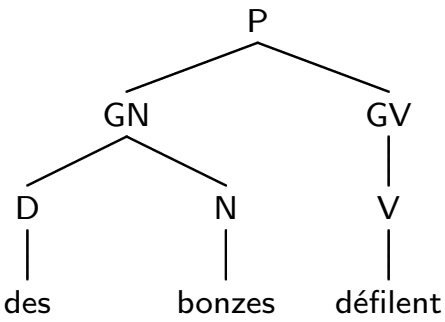
des bonzes défilent pacifiquement $\longrightarrow$



2. Génération de sous-arbres prétendants



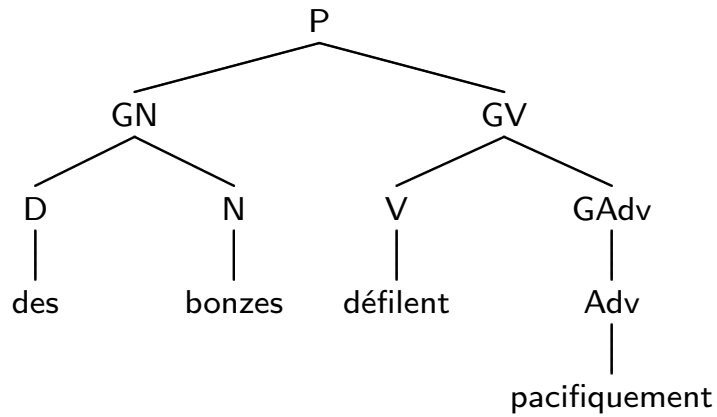




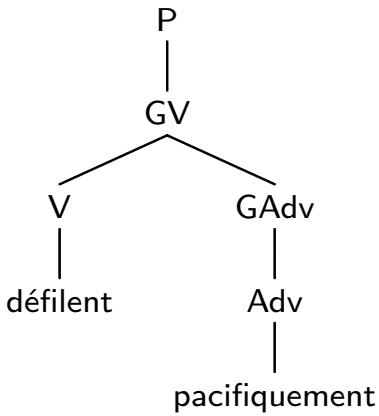
[Knight & Marcu, 2002] : un modèle de canal bruité

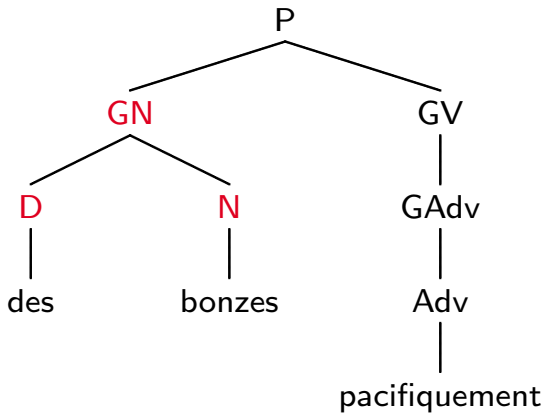
1. Analyse syntaxique, arbres en constituants [Collins, 1997]

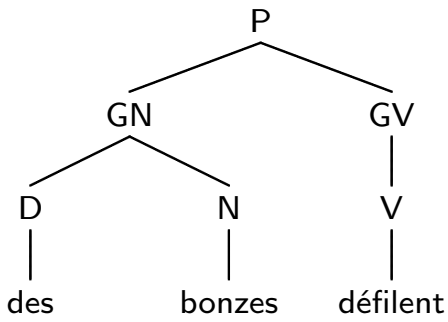
des bonzes défilent pacifiquement $\longrightarrow$



2. Génération de sous-arbres prétendants



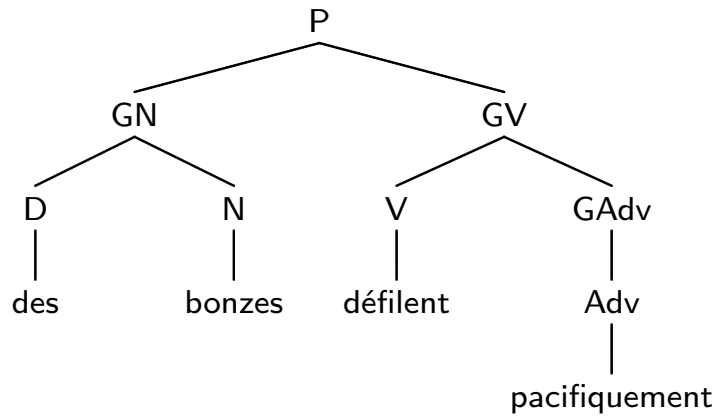




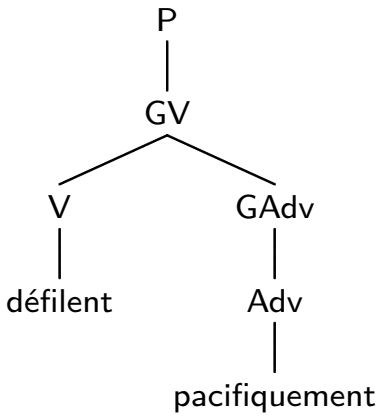
[Knight & Marcu, 2002] : un modèle de canal bruité

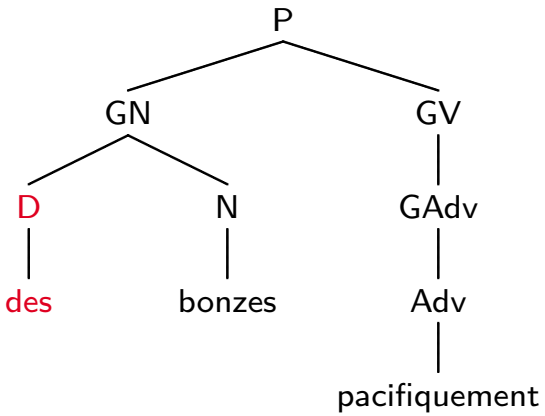
1. Analyse syntaxique, arbres en constituants [Collins, 1997]

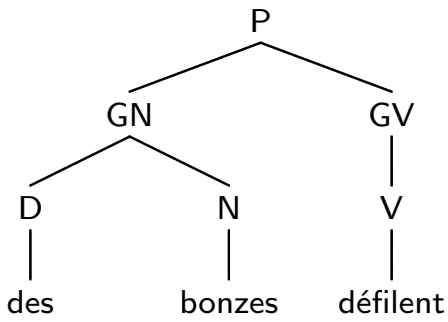
des bonzes défilent pacifiquement $\longrightarrow$



2. Génération de sous-arbres prétendants



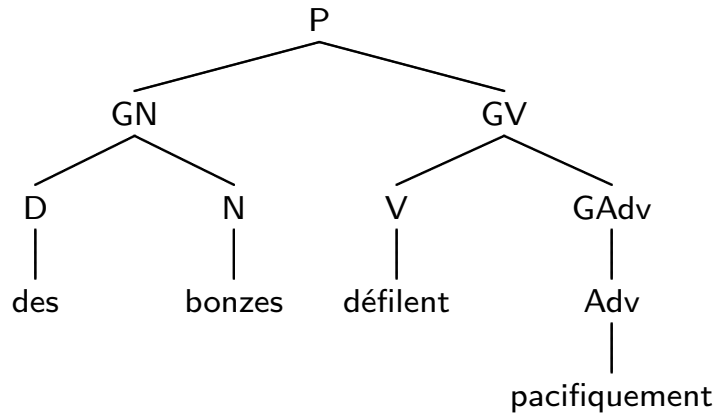




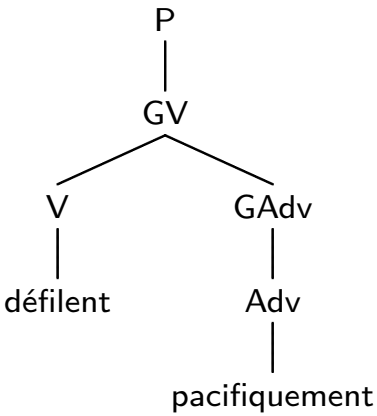
[Knight & Marcu, 2002] : un modèle de canal bruité

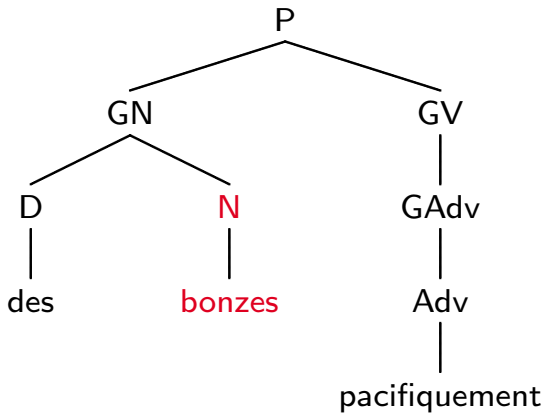
1. Analyse syntaxique, arbres en constituants [Collins, 1997]

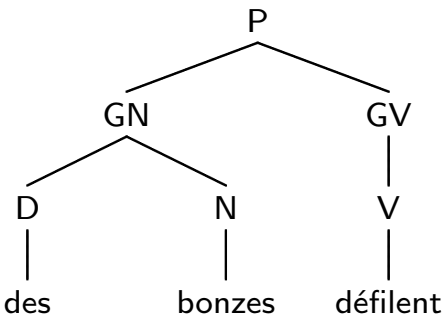
des bonzes défilent pacifiquement $\longrightarrow$



2. Génération de sous-arbres prétendants



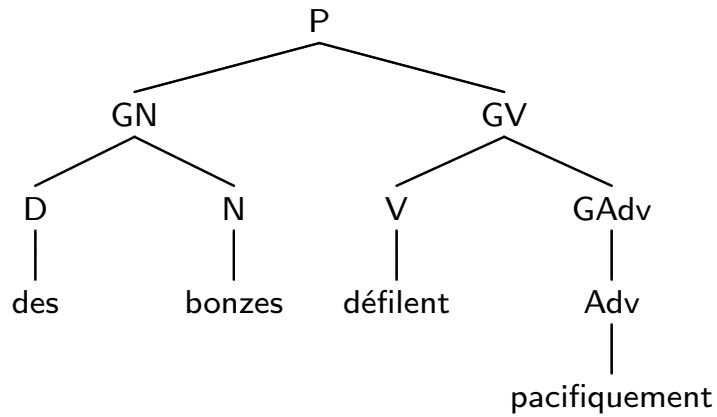




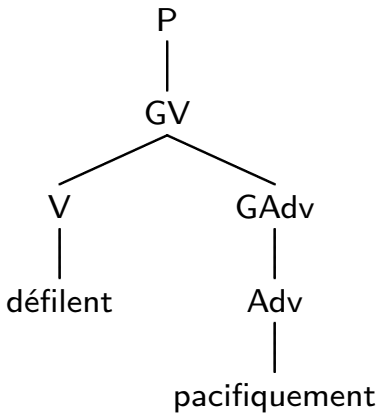
[Knight & Marcu, 2002] : un modèle de canal bruité

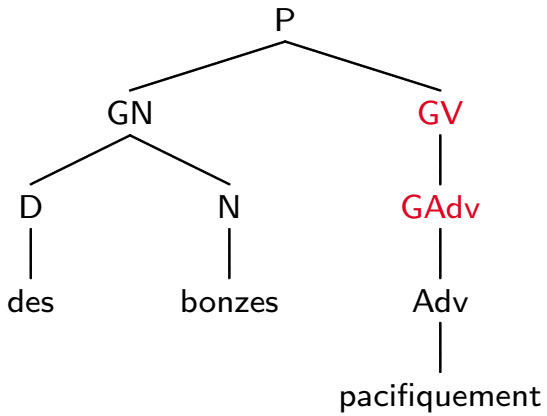
1. Analyse syntaxique, arbres en constituants [Collins, 1997]

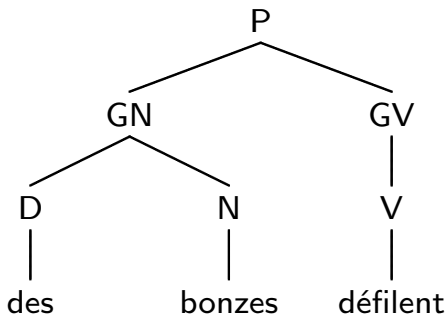
des bonzes défilent pacifiquement $\longrightarrow$



2. Génération de sous-arbres prétendants



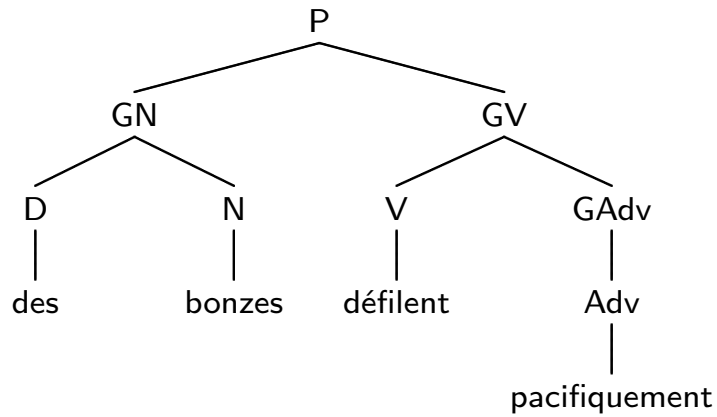




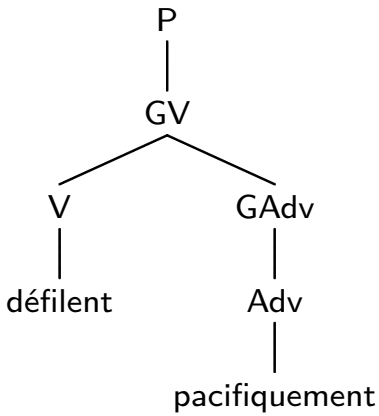
[Knight & Marcu, 2002] : un modèle de canal bruité

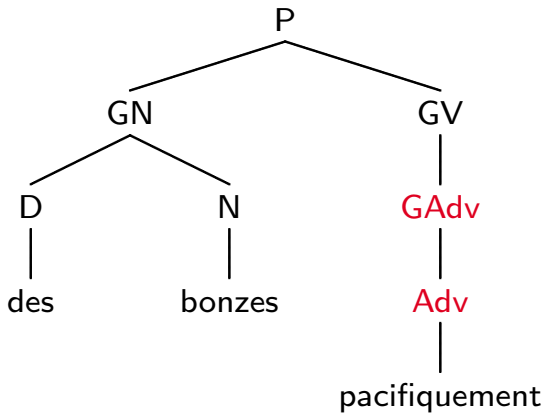
1. Analyse syntaxique, arbres en constituants [Collins, 1997]

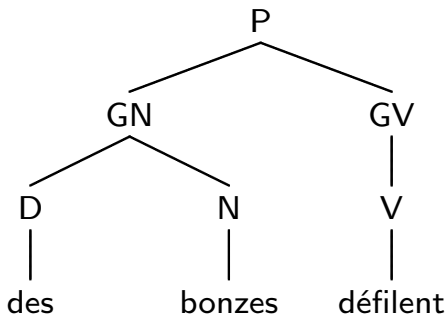
des bonzes défilent pacifiquement $\longrightarrow$



2. Génération de sous-arbres prétendants



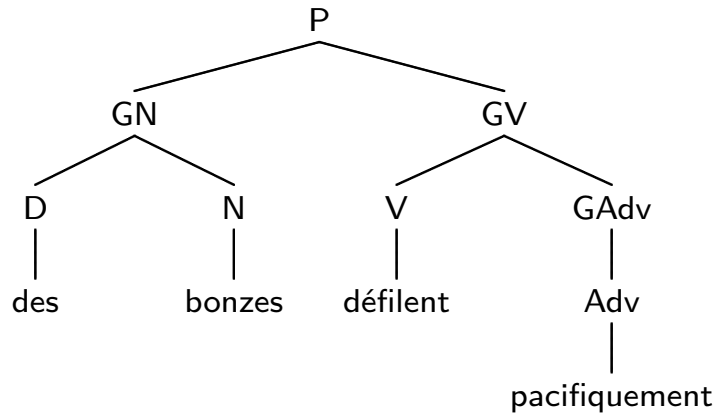




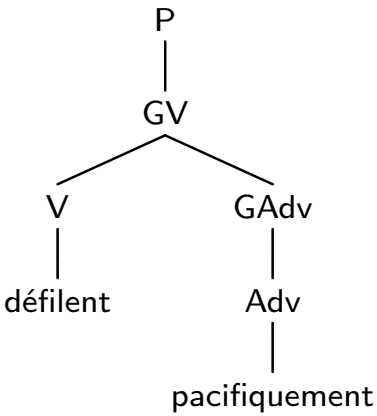
[Knight & Marcu, 2002] : un modèle de canal bruité

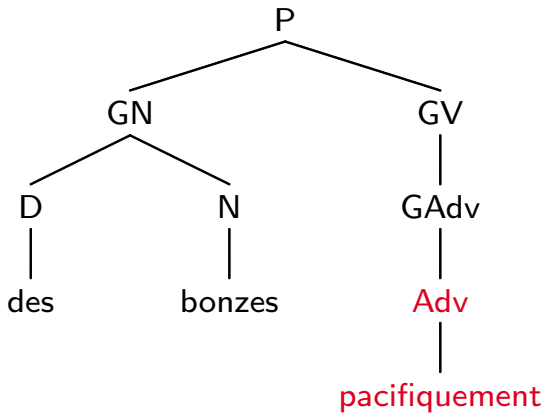
1. Analyse syntaxique, arbres en constituants [Collins, 1997]

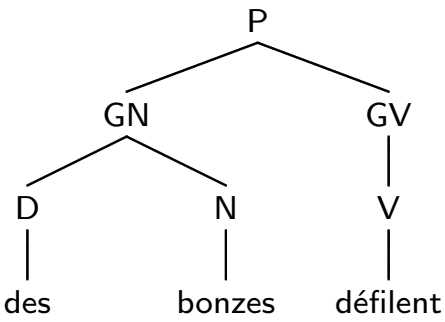
des bonzes défilent pacifiquement $\longrightarrow$



2. Génération de sous-arbres prétendants







[Knight & Marcu, 2002] : un modèle de canal bruité

3. par phrase, choix de l'arbre le plus probable
   - entraînement : corpus (Ziff–Davis) de documents d'actualités associés à leur résumé
   - par nœud, score de Grammaires Hors-Contexte Probabilistes

   $\Rightarrow$ grammaticalité, ex : $P_{GHC}(P \rightarrow GN\ GV|P)$, $P_{GHC}(V \rightarrow \textit{défilent}|V)$

[Knight & Marcu, 2002] : un modèle de canal bruité

3 par phrase, choix de l'arbre le plus probable

- entraînement : corpus (Ziff–Davis) de documents d'actualités associés à leur résumé
- par nœud, score de Grammaires Hors-Contexte Probabilistes

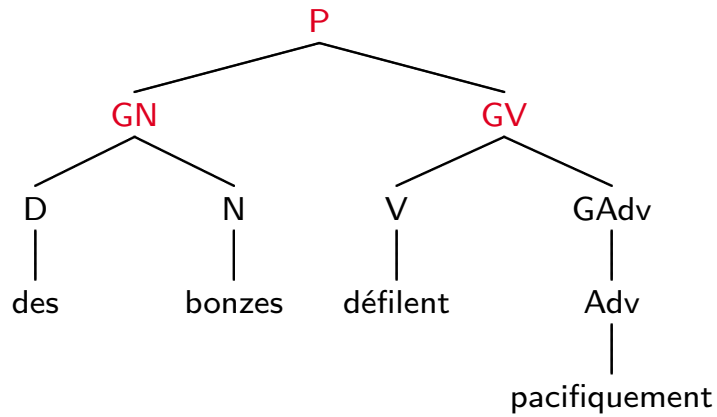
  $\Rightarrow$ grammaticalité, ex : $P_{GHC}(P \rightarrow GN\ GV|P)$, $P_{GHC}(V \rightarrow \textit{défilent}|V)$

  $\Rightarrow$ expansion, ex : $P_{exp}(P \rightarrow GN\ GV|P \rightarrow GN\ GV)$, $P_{exp}(GV \rightarrow V|GV \rightarrow V\ GAdv)$

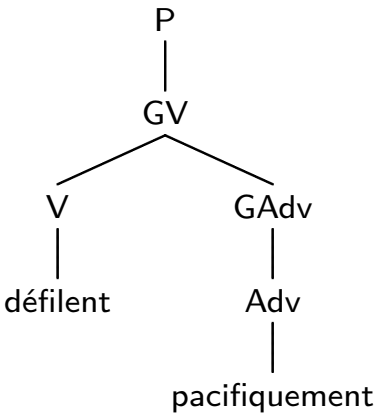
[Knight & Marcu, 2002] : un modèle de canal bruité

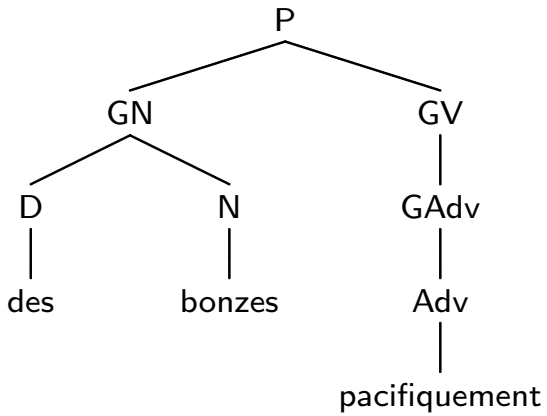
1. Analyse syntaxique, arbres en constituants [Collins, 1997]

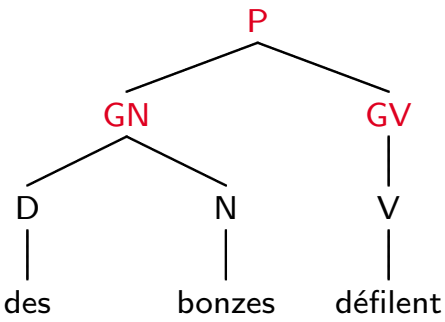
des bonzes défilent pacifiquement $\longrightarrow$



2. Génération de sous-arbres prétendants



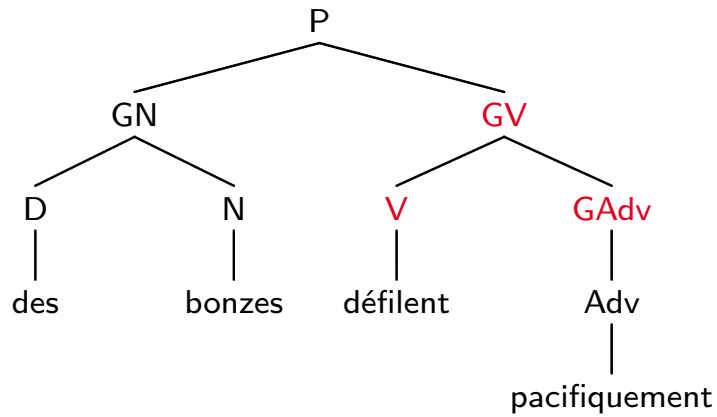




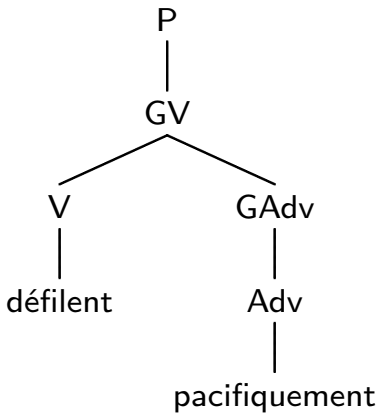
[Knight & Marcu, 2002] : un modèle de canal bruité

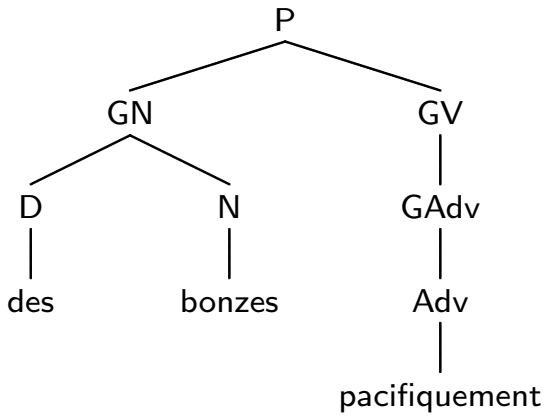
1. Analyse syntaxique, arbres en constituants [Collins, 1997]

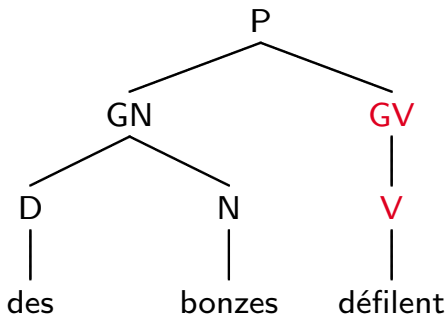
des bonzes défilent pacifiquement $\longrightarrow$



2. Génération de sous-arbres prétendants







[Knight & Marcu, 2002] : un modèle de canal bruité

3 par phrase, choix de l'arbre le plus probable

- entraînement : corpus (Ziff–Davis) de documents d'actualités associés à leur résumé
- par nœud, score de Grammaires Hors-Contexte Probabilistes

  $\Rightarrow$ grammaticalité, ex : $P_{GHC}(P \rightarrow GN\ GV|P)$, $P_{GHC}(V \rightarrow \textit{défilent}|V)$

  $\Rightarrow$ expansion, ex : $P_{exp}(P \rightarrow GN\ GV|P \rightarrow GN\ GV)$, $P_{exp}(GV \rightarrow V|GV \rightarrow V\ GAdv)$

[Knight & Marcu, 2002] : un modèle de canal bruité

3 par phrase, choix de l'arbre le plus probable

- entraînement : corpus (Ziff–Davis) de documents d'actualités associés à leur résumé
- par nœud, score de Grammaires Hors-Contexte Probabilistes

  $\Rightarrow$ grammaticalité, ex : $P_{GHC}(P \rightarrow GN\ GV|P)$, $P_{GHC}(V \rightarrow \textit{défilent}|V)$

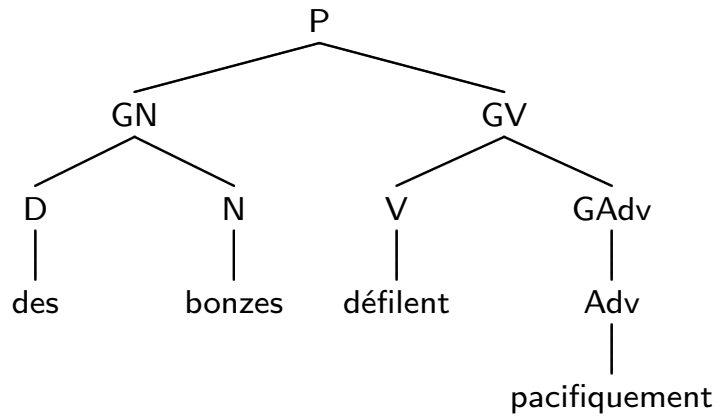
  $\Rightarrow$ expansion, ex : $P_{exp}(P \rightarrow GN\ GV|P \rightarrow GN\ GV)$, $P_{exp}(GV \rightarrow V|GV \rightarrow V\ GAdv)$
- par couple de feuilles, score de bi-grammes de mots

  $\Rightarrow$ ex : $P_{bigram}(des|\emptyset)$, $P_{bigram}(bonzes|des)$, $P_{bigram}(pacifiquement|bonzes)$

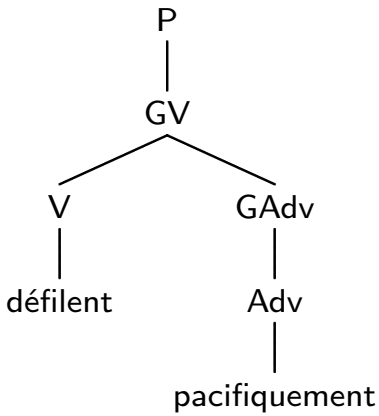
[Knight & Marcu, 2002] : un modèle de canal bruité

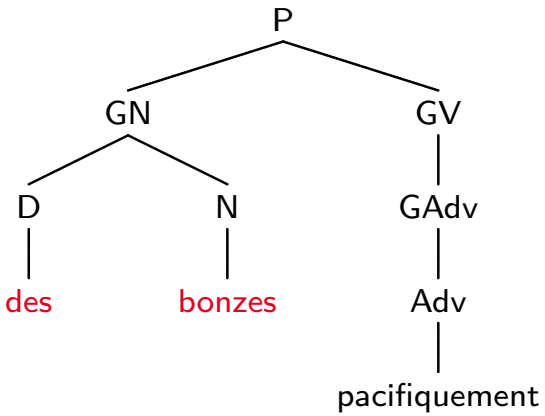
1. Analyse syntaxique, arbres en constituants [Collins, 1997]

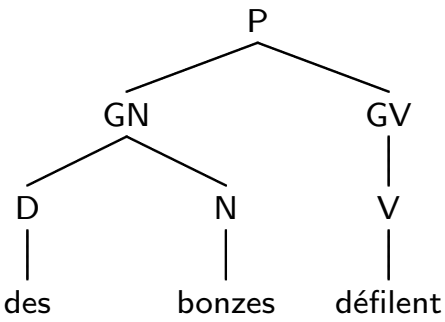
des bonzes défilent pacifiquement $\longrightarrow$



2. Génération de sous-arbres prétendants



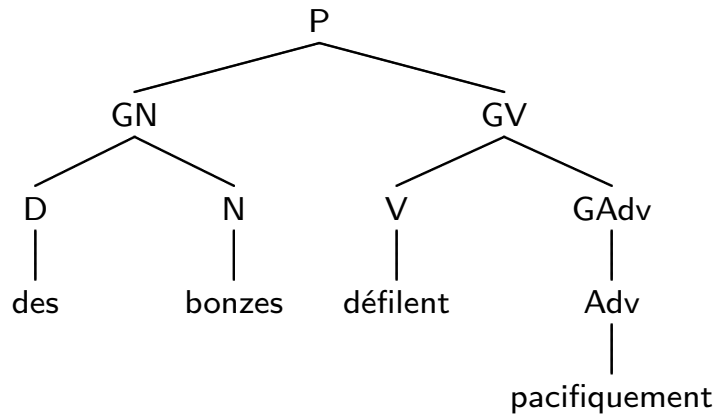




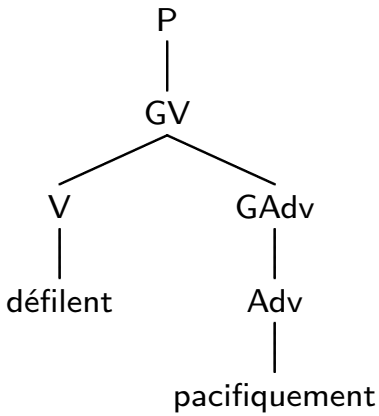
[Knight & Marcu, 2002] : un modèle de canal bruité

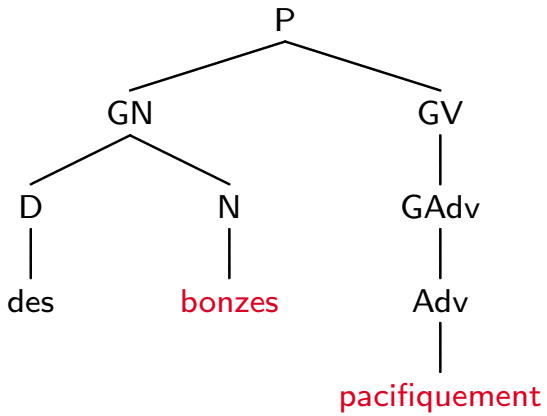
1. Analyse syntaxique, arbres en constituants [Collins, 1997]

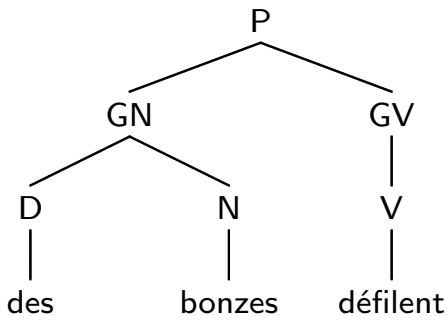
des bonzes défilent pacifiquement $\longrightarrow$



2. Génération de sous-arbres prétendants







[Knight & Marcu, 2002] : un modèle de canal bruité

3 par phrase, choix de l'arbre le plus probable

- entraînement : corpus (Ziff–Davis) de documents d'actualités associés à leur résumé
- par nœud, score de Grammaires Hors-Contexte Probabilistes

  $\Rightarrow$ grammaticalité, ex : $P_{GHC}(P \rightarrow GN\ GV|P)$, $P_{GHC}(V \rightarrow \textit{défilent}|V)$

  $\Rightarrow$ expansion, ex : $P_{exp}(P \rightarrow GN\ GV|P \rightarrow GN\ GV)$, $P_{exp}(GV \rightarrow V|GV \rightarrow V\ GAdv)$
- par couple de feuilles, score de bi-grammes de mots

  $\Rightarrow$ ex : $P_{bigram}(des|\emptyset)$, $P_{bigram}(bonzes|des)$, $P_{bigram}(pacifiquement|bonzes)$

[Knight & Marcu, 2002] : un modèle de canal bruité

3 par phrase, choix de l'arbre le plus probable

- entraînement : corpus (Ziff–Davis) de documents d'actualités associés à leur résumé
- par nœud, score de Grammaires Hors-Contexte Probabilistes

  $\Rightarrow$ grammaticalité, ex : $P_{GHC}(P \rightarrow GN\ GV|P)$, $P_{GHC}(V \rightarrow \textit{défilent}|V)$

  $\Rightarrow$ expansion, ex : $P_{exp}(P \rightarrow GN\ GV|P \rightarrow GN\ GV)$, $P_{exp}(GV \rightarrow V|GV \rightarrow V\ GAdv)$
- par couple de feuilles, score de bi-grammes de mots

  $\Rightarrow$ ex : $P_{bigram}(des|\emptyset)$, $P_{bigram}(bonzes|des)$, $P_{bigram}(pacifiquement|bonzes)$
- par taille d'arbre, conservation du mieux noté

[Knight & Marcu, 2002] : un modèle de canal bruité

3 par phrase, choix de l'arbre le plus probable

- entraînement : corpus (Ziff–Davis) de documents d'actualités associés à leur résumé
- par nœud, score de Grammaires Hors-Contexte Probabilistes

  $\Rightarrow$ grammaticalité, ex : $P_{GHC}(P \rightarrow GN\ GV|P)$, $P_{GHC}(V \rightarrow \textit{défilent}|V)$

  $\Rightarrow$ expansion, ex : $P_{exp}(P \rightarrow GN\ GV|P \rightarrow GN\ GV)$, $P_{exp}(GV \rightarrow V|GV \rightarrow V\ GAdv)$
- par couple de feuilles, score de bi-grammes de mots

  $\Rightarrow$ ex : $P_{bigram}(des|\emptyset)$, $P_{bigram}(bonzes|des)$, $P_{bigram}(pacifiquement|bonzes)$
- par taille d'arbre, conservation du mieux noté
- choix final : quotient entre probabilité et nombre de feuilles

[Knight & Marcu, 2002] : un modèle de canal bruité

4. par arbre choisi, reconstitution de la phrase

[Knight & Marcu, 2002] : un modèle de canal bruité

4. par arbre choisi, reconstitution de la phrase

5. reconstitution du texte (concaténation des phrases)

[Knight & Marcu, 2002] : un modèle de canal bruité

4. par arbre choisi, reconstitution de la phrase

5. reconstitution du texte (concaténation des phrases)

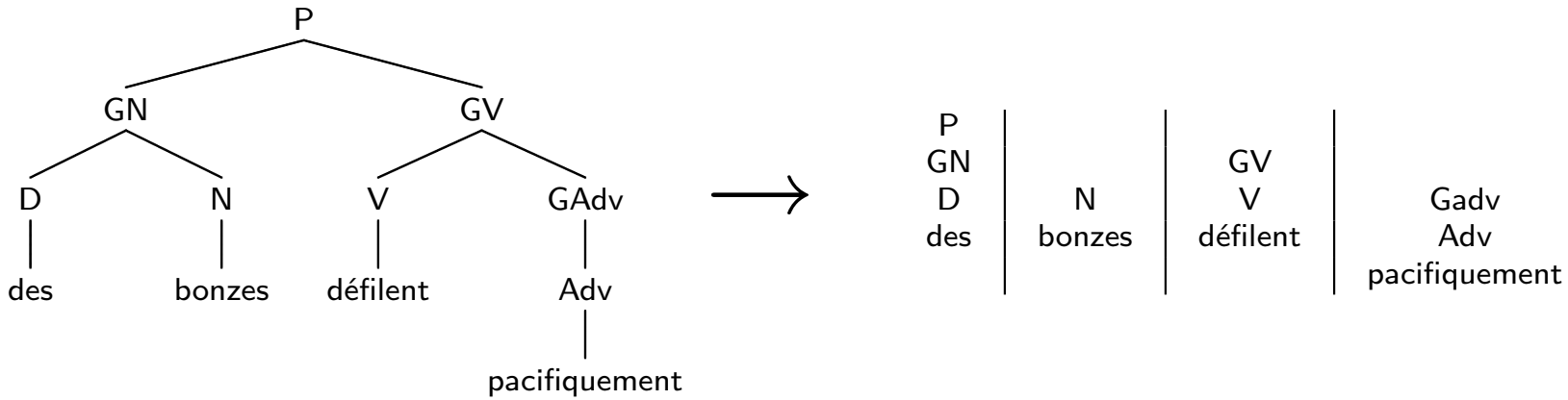
[Knight & Marcu, 2002] : un modèle fondé sur la décision

1. Analyse syntaxique

[Knight & Marcu, 2002] : un modèle fondé sur la décision

1. Analyse syntaxique

2. Décomposition de l'arbre en une liste d'entrée à leur algorithme



[Knight & Marcu, 2002] : un modèle fondé sur la décision

3 Reconstruction d'un nouvel arbre

[Knight & Marcu, 2002] : un modèle fondé sur la décision

3 Reconstruction d'un nouvel arbre
$\Rightarrow$ 4 opérations : SHIFT, REDUCE, DROP et ASSIGN

[Knight & Marcu, 2002] : un modèle fondé sur la décision

3 Reconstruction d'un nouvel arbre

$\Rightarrow$ 4 opérations : SHIFT, REDUCE, DROP et ASSIGN

pile	**liste d'entrée**				**opérations**
	P GN D des	N bonzes	GV V défilent	Gadv Adv pacifiquement	SHIFT, ASSIGN D

[Knight & Marcu, 2002] : un modèle fondé sur la décision

3 Reconstruction d'un nouvel arbre

$\Rightarrow$ 4 opérations : SHIFT, REDUCE, DROP et ASSIGN

pile	liste d'entrée				opérations
	P GN D des	N bonzes	GV V défilent	Gadv Adv pacifiquement	SHIFT, ASSIGN D
D | des		N bonzes	GV V défilent	Gadv Adv pacifiquement	SHIFT, ASSIGN N

[Knight & Marcu, 2002] : un modèle fondé sur la décision

3 Reconstruction d'un nouvel arbre

⇒ 4 opérations : SHIFT, REDUCE, DROP et ASSIGN

pile	liste d'entrée				opérations
	P GN D des	N bonzes	GV V défilent	Gadv Adv pacifiquement	SHIFT, ASSIGN D
D | des		N bonzes	GV V défilent	Gadv Adv pacifiquement	SHIFT, ASSIGN N
D N | | des bonzes			GV V défilent	Gadv Adv pacifiquement	REDUCE 2 GN

[Knight & Marcu, 2002] : un modèle fondé sur la décision

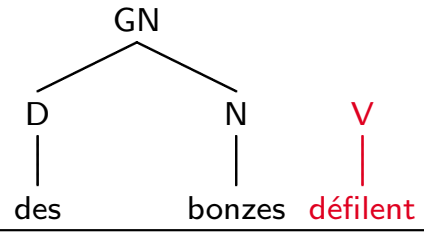
3 Reconstruction d'un nouvel arbre

⇒ 4 opérations : SHIFT, REDUCE, DROP et ASSIGN

pile	liste d'entrée				opérations
	P GN D des	N bonzes	GV V défilent	Gadv Adv pacifiquement	SHIFT, ASSIGN D
D | des		N bonzes	GV V défilent	Gadv Adv pacifiquement	SHIFT, ASSIGN N
D N | | des bonzes			GV V défilent	Gadv Adv pacifiquement	REDUCE 2 GN
GN D N | | des bonzes			GV V défilent	Gadv Adv pacifiquement	SHIFT, ASSIGN V

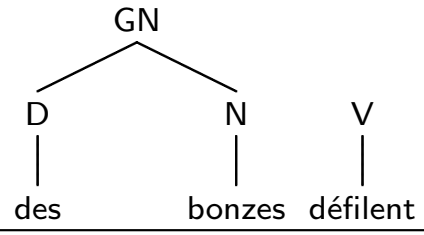
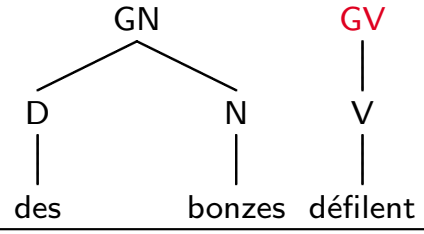
[Knight & Marcu, 2002] : un modèle fondé sur la décision

3 Reconstruction d'un nouvel arbre (suite)

pile	liste d'entrée				opérations
				Gadv Adv pacifiquement	REDUCE 1 GV

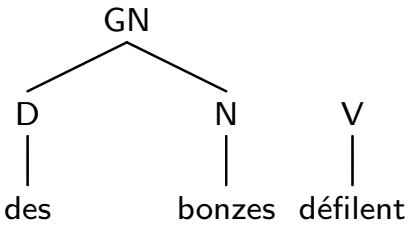
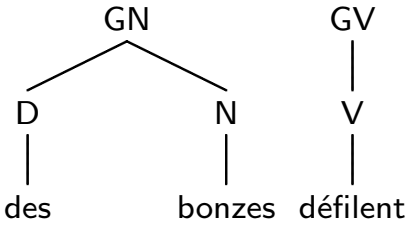
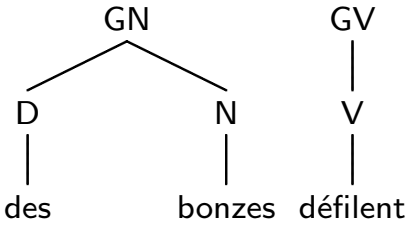
[Knight & Marcu, 2002] : un modèle fondé sur la décision

3 Reconstruction d'un nouvel arbre (suite)

pile	liste d'entrée				opérations
				Gadv Adv pacifiquement	REDUCE 1 GV
				Gadv Adv pacifiquement	DROP GAdv

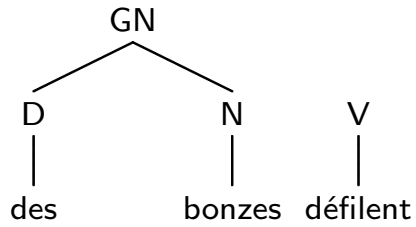
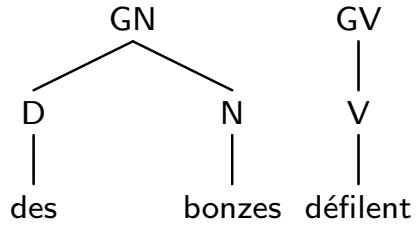
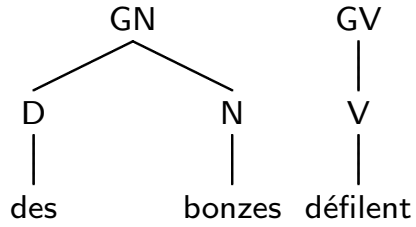
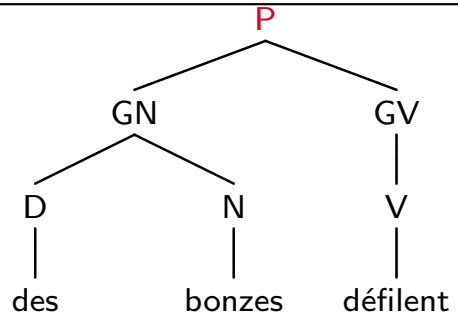
[Knight & Marcu, 2002] : un modèle fondé sur la décision

3 Reconstruction d'un nouvel arbre (suite)

pile	liste d'entrée				opérations
				Gadv Adv pacifiquement	REDUCE 1 GV
				Gadv Adv pacifiquement	DROP GAdv
					REDUCE 2 P

[Knight & Marcu, 2002] : un modèle fondé sur la décision

3 Reconstruction d'un nouvel arbre (suite)

pile	liste d'entrée				opérations
				Gadv Adv pacifiquement	REDUCE 1 GV
				Gadv Adv pacifiquement	DROP GAdv
					REDUCE 2 P
					

[Knight & Marcu, 2002] : un modèle fondé sur la décision

3. Reconstruction d'un nouvel arbre (suite)
   - moteur d'apprentissage[7] pour décider quelle opération effectuer

[7] programme C4.5

[Knight & Marcu, 2002] : un modèle fondé sur la décision

3. Reconstruction d'un nouvel arbre (suite)
   - moteur d'apprentissage[7] pour décider quelle opération effectuer
     ⇒ entraînement : corpus Ziff–Davis
     ⇒ selon la configuration de l'arbre (la pile)
     ⇒ selon le contenu de la liste d'entrée
     ⇒ selon la dernière opération effectuée

[7] programme C4.5

[Knight & Marcu, 2002] : un modèle fondé sur la décision

Différences avec le modèle de canal bruité

- davantage de transformations
  - ✓ espace des arbres générables plus grand
  - ✗ grammaticalité inférieure

[Knight & Marcu, 2002] : un modèle fondé sur la décision

Différences avec le modèle de canal bruité

- davantage de transformations
  - ✓ espace des arbres générables plus grand
  - ✗ grammaticalité inférieure
- mots exclus de l'aglorithme
  - ✓ pas de contrainte sur le vocabulaire

[Knight & Marcu, 2002] : un modèle fondé sur la décision

Différences avec le modèle de canal bruité

- davantage de transformations
  - ✓ espace des arbres générables plus grand
  - ✗ grammaticalité inférieure
- mots exclus de l'aglorithme
  - ✓ pas de contrainte sur le vocabulaire
- algorithme déterministe
  - ✓ compression rapide
  - ✗ un seul taux de compression possible

[Hovy *et al.*, 2005], *Basic Elements*

1. Génération de *Basic Elements* et analyse syntaxique

[Hovy *et al.*, 2005], *Basic Elements*

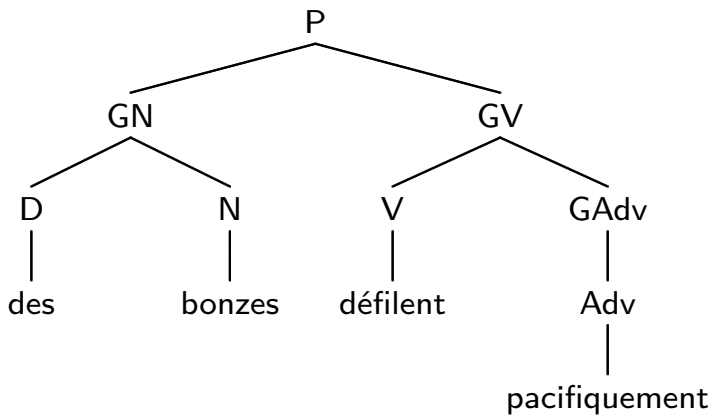
1 Génération de *Basic Elements* et analyse syntaxique

des bonzes défilent pacifiquement

***Basic Elements* (types de dépendances)**

tête	dépendant	relation
bonzes	des	déterminant
défilent	bonzes	sujet
défilent	pacifiquement	adverbe

analyse syntaxique [Collins, 1997]



[Hovy *et al.*, 2005], *Basic Elements*

2 classement des BE par importance
$\Rightarrow$ score de rapport de vraisemblance[8]

[8] importance relative aux BE du corpus

[Hovy *et al.*, 2005], *Basic Elements*

2. classement des BE par importance
   $\Rightarrow$ score de rapport de vraisemblance[8]

3. seuil d'importance arbitraire
   $\Rightarrow$ 2 classes de BE (importants ou pas)

[8] importance relative aux BE du corpus

[Hovy *et al.*, 2005], *Basic Elements*

2. classement des BE par importance
   $\Rightarrow$ score de rapport de vraisemblance[8]

3. seuil d'importance arbitraire
   $\Rightarrow$ 2 classes de BE (importants ou pas)

4. génération d'arbres élagués pour chaque dépendant de chaque BE peu important

[8] importance relative aux BE du corpus

[Hovy *et al.*, 2005], *Basic Elements*

2. classement des BE par importance
   $\Rightarrow$ score de rapport de vraisemblance[8]

3. seuil d'importance arbitraire
   $\Rightarrow$ 2 classes de BE (importants ou pas)

4. génération d'arbres élagués pour chaque dépendant de chaque BE peu important
   $\Rightarrow$ ex : « pacifiquement » peu important dans BE(défilent | pacifiquement | adverbe)

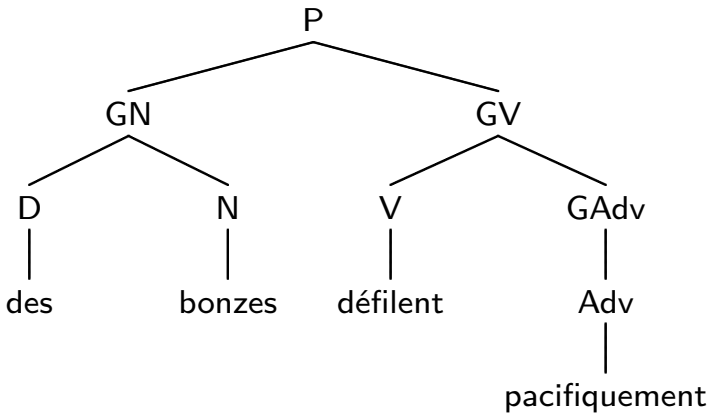
[8] importance relative aux BE du corpus

[Hovy *et al.*, 2005], *Basic Elements*

2. classement des BE par importance
   ⇒ score de rapport de vraisemblance[8]

3. seuil d'importance arbitraire
   ⇒ 2 classes de BE (importants ou pas)

4. génération d'arbres élagués pour chaque dépendant de chaque BE peu important
   ⇒ ex : « pacifiquement » peu important dans BE(défilent | pacifiquement | adverbe)



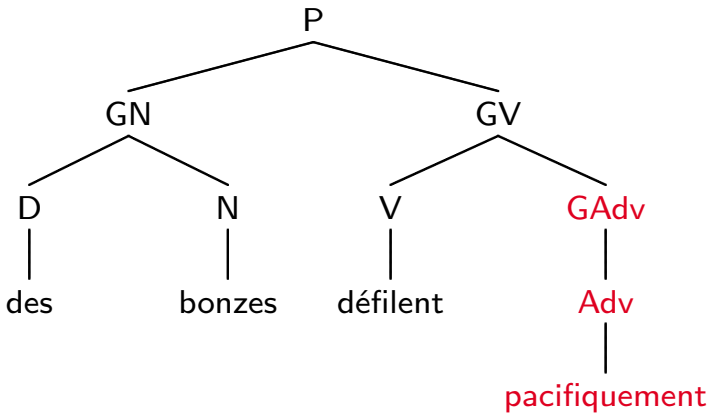
[8] importance relative aux BE du corpus

[Hovy *et al.*, 2005], *Basic Elements*

2. classement des BE par importance
   ⇒ score de rapport de vraisemblance[8]

3. seuil d'importance arbitraire
   ⇒ 2 classes de BE (importants ou pas)

4. génération d'arbres élagués pour chaque dépendant de chaque BE peu important
   ⇒ ex : « pacifiquement » peu important dans BE(défilent | pacifiquement | adverbe)



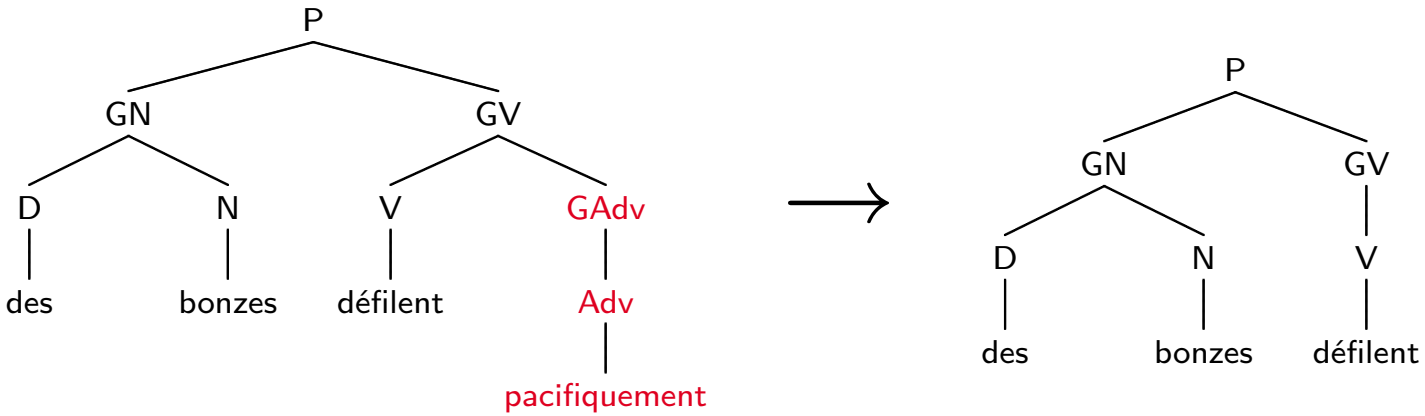
[8] importance relative aux BE du corpus

[Hovy *et al.*, 2005], *Basic Elements*

2. classement des BE par importance
   ⇒ score de rapport de vraisemblance[8]

3. seuil d'importance arbitraire
   ⇒ 2 classes de BE (importants ou pas)

4. génération d'arbres élagués pour chaque dépendant de chaque BE peu important
   ⇒ ex : « pacifiquement » peu important dans BE(défilent | pacifiquement | adverbe)



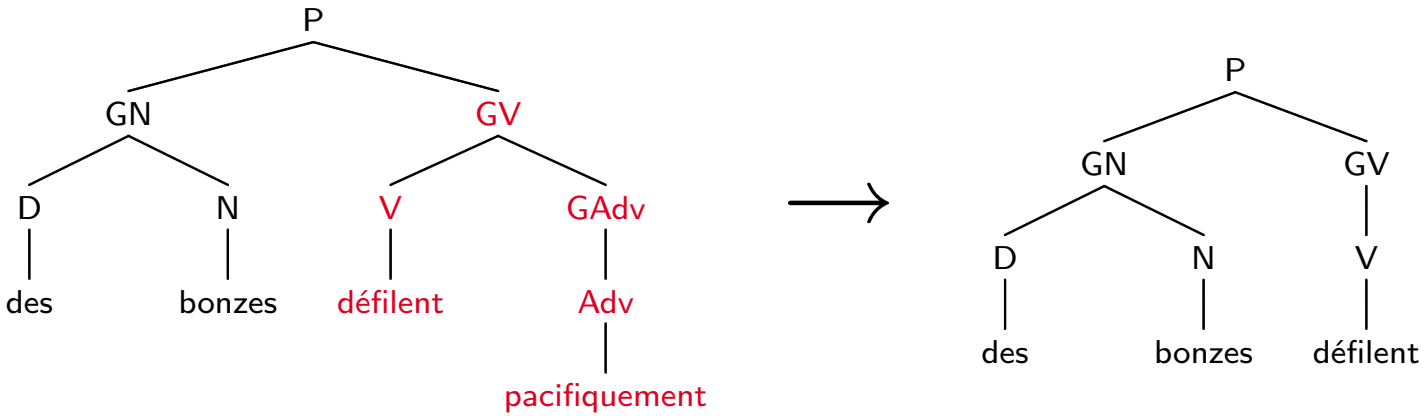
[8] importance relative aux BE du corpus

[Hovy *et al.*, 2005], *Basic Elements*

2. classement des BE par importance
   $\Rightarrow$ score de rapport de vraisemblance[8]

3. seuil d'importance arbitraire
   $\Rightarrow$ 2 classes de BE (importants ou pas)

4. génération d'arbres élagués pour chaque dépendant de chaque BE peu important
   $\Rightarrow$ ex : « pacifiquement » peu important dans BE(défilent | pacifiquement | adverbe)



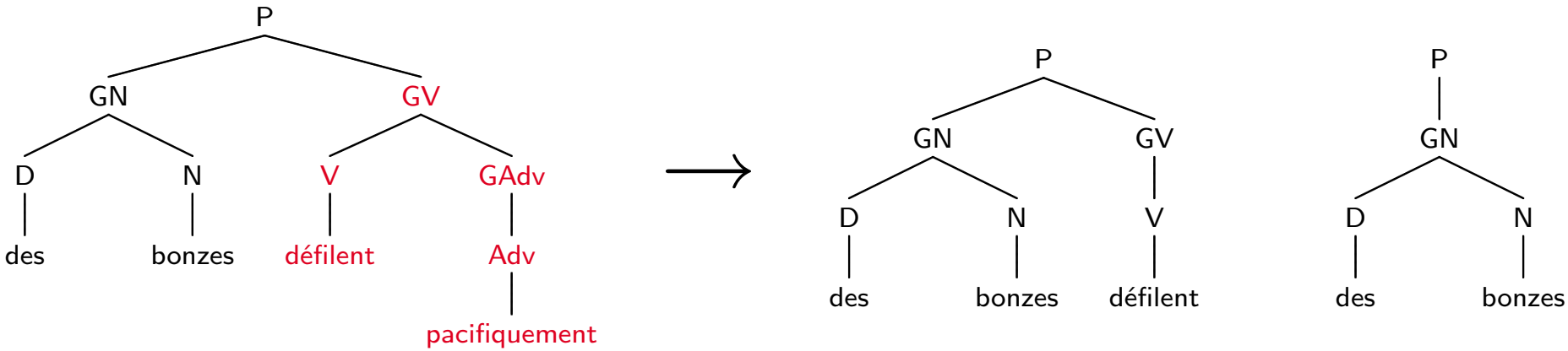
[8] importance relative aux BE du corpus

[Hovy *et al.*, 2005], *Basic Elements*

2. classement des BE par importance
   ⇒ score de rapport de vraisemblance[8]

3. seuil d'importance arbitraire
   ⇒ 2 classes de BE (importants ou pas)

4. génération d'arbres élagués pour chaque dépendant de chaque BE peu important
   ⇒ ex : « pacifiquement » peu important dans BE(défilent | pacifiquement | adverbe)



[8] importance relative aux BE du corpus

[Hovy *et al.*, 2005], *Basic Elements*

5 par arbre élagué, choix du plus probable syntaxiquement

⇒ score de Grammaires Hors-Contexte Probabilistes

⇒ entraînement : corpus étiqueté (Penn TreeBank)

⇒ prise en compte de BE importants élagués

[Hovy *et al.*, 2005], *Basic Elements*

5 par arbre élagué, choix du plus probable syntaxiquement

⇒ score de Grammaires Hors-Contexte Probabilistes

⇒ entraînement : corpus étiqueté (Penn TreeBank)

⇒ prise en compte de BE importants élagués

6 par arbre choisi, reconstitution de la phrase

[Hovy *et al.*, 2005], *Basic Elements*

5. par arbre élagué, choix du plus probable syntaxiquement

   ⇒ score de Grammaires Hors-Contexte Probabilistes

   ⇒ entraînement : corpus étiqueté (Penn TreeBank)

   ⇒ prise en compte de BE importants élagués

6. par arbre choisi, reconstitution de la phrase

7. reconstitution du texte (concaténation des phrases)