Les discussions Wikipedia : un corpus pour caractériser le genre "discussion"

Lydia-Mai Ho-Dac et Véronika Laippala

CLLE-ERSS, TIAS

CMO, 24-25 octobre, Rennes

Plan

1. Motivations : Web As Corpus
2. Wikipedia As Corpus
3. Premières analyses
4. Conclusions et perspectives

Plan

1. Motivations : Web As Corpus
2. Wikipedia As Corpus
3. Premières analyses
4. Conclusions et perspectives

Accès facile, des données disparates

- Plusieurs corpus automatiquement collectés du Web
  - WaC (Baroni & al. 2009), Finnish Internet Parsebank (FIP, Kanerva & al. 2014), etc.
- Avantages : Taille importante, accès facile*
- Inconvénients : Contenu très hétérogène compliquant le traitement des données
  - Traductions (semi)automatiques fréquentes ($\tilde{1}$0% dans le FIP)
  - Grande variété de genres et registres, dont certains inconnus

*mais pas nécessairement distribué publiquement

A Corpus Factory for Many Languages

(Kilgarriff et al. 2010)

A	Arabic (Arabic web corpus)
B	Basque (basque_WaC) Bengali (bengaliWaC) Bosnian (bosnianWaC14)
C	Cantonese (Cantonese WaC) Chinese (ChineseTaiwanWaC) Croatian (hrWaC, hrWaC_10M)
D	Danish (danishWaC) Dutch (Dutch web corpus, nlWaC, nlWaC_1)
E	English (pukWaC, ukWaC, ukWaC_1, ukWaC_10M, ukWaC_10M_1, ukWaC2, ukWaC2_1, ukWaC3, ukWaC_mcd, ukWaCsst)
F	Filipino (filipinoWaC) Finnish (finnishWaC) Frisian (frisianWaC) French (frWaC, frWaC1_1)
G	Georgian (georgianWaC) German (deWaC, Parsed DeWaC (sDeWaC)) Greek (gkWaC) Gujarati (gujarathiWaC)
H	Hebrew (hebWaC) Hindi (hindiWaC, hindiWaC3)
I	Igbo (igboWaC) Indonesian (indonesianWaC) Italian (itWaC)
J	Japanese (jpWaC, jpWaC_10M, jpWaC2)
K	Korean (koreanWaC) Kannada (Kannada WaC)
L	Latin (latinWaC, latinWaC2) Latvian (latvianWaC, latvianWaC_shallow) Lithuanian (lithuanianWaC, lithuanianWaC_v2, lithuanianWaC_v2_10M)
M	Malay (malayalamWaC, malaysianWaC2) Maltese (malteseWaC, malteseWaC2, malteseWaC2_sample) Maori (maoriWaC)
N	Nepali (nepaliWaC) Norwegian (norwegianWaC)
P	Persian (WBC-Per) Polish (Polish Web Corpus)
R	Romanian (romanian_WaC) Russian (Russian Web Corpus)
S	Samoan (SamoanWaC) Serbian (serbianWaC, serbianWaC14, srWaC, srWaC22M) Setswana (setswanaWaC, setswanaWaC2) Spanish (Spanish wen corpus) Swahili (swahiliWaC, swahiliWaC_1) Swedish (swedishWaC, swedish_WaC, swedish_WaC_10M)
T	Tamil (tamilWaC) Tatar (Tatar Sample) Telugu (teluguWaC, teluguWaC2) Thai (thaiWaC) Turkish (turkishWaC, turkishWaC2, turkishWaC2_1, turkishWaC2_1_s, turkishWaC2_1_uniattr)
U	Urdu
V	vietnameseWaC2 (Viatnamese)
W	Welsh (welshWaC)
Y	Yoruba (Yoruba web corpus)

Enjeux : une large variété de "genres"

La notion de genre

By means the concept of genre we can approach texts from the macro-level as communicative acts within a discourse network or system (Trosborg 1997 :7)

le genre est une « catégorie de textes fondée sur une pratique sociale établie, définie a priori. La catégorie est reconnue et validée par le fait qu'elle peut se dénommer. » (Gayral et al. 2007 :6)

Enjeux : définir les *genres* du web

- Nécessité de profiler les textes
- Nécessite de comprendre les proportions des genres / registres différents

$\Rightarrow$ Développer des méthodes quantitatives pour l'analyse et

Plan

1. Motivations : Web As Corpus
2. Wikipedia As Corpus
   - Les discussions Wikipedia
   - Constitution du corpus WikiDiscussion
3. Premières analyses
4. Conclusions et perspectives

Wikipedia As Corpus

Accessibilité et quantité des données

- contenu libre distribué publiquement (Creative Commons by-sa)
- depuis 2001
- existe dans presque toutes les langues $\rightarrow$ objet d'étude international

Au 11 mai 2015 : 1 622 066 articles, 366 326 discussions associées aux articles, 16 192 contributeurs (Wikipédiens) ayant fait au moins une modification ces 30 derniers jours, 5 000 qui en ont fait au moins 5 et près de 800 qui en ont fait au moins 100.

données fournies par le projet français Wikipedia

Wikipedia As Corpus

Accessibilité et quantité des données

- contenu libre distribué publiquement (Creative Commons by-sa)
- depuis 2001
- existe dans presque toutes les langues $\rightarrow$ objet d'étude international

Variété de genres et de situations de communication

- articles encyclopédiques
- discussions

Wikipedia As Corpus

Accessibilité et quantité des données

- contenu libre distribué publiquement (Creative Commons by-sa)
- depuis 2001
- existe dans presque toutes les langues → objet d'étude international

Variété de genres et de situations de communication

- articles encyclopédiques
- **discussions**
  - autour de la rédaction collaborative d'un article
  - «cafés et bistrots» («Forum des Nouveaux», «Le salon de médiation», «Legifer» ...)

Wikipedia As Corpus

Accessibilité et quantité des données

- contenu libre distribué publiquement (Creative Commons by-sa)
- depuis 2001
- existe dans presque toutes les langues → objet d'étude international

Variété de genres et de situations de communication

- articles encyclopédiques
- **discussions**
  - autour de la rédaction collaborative d'un article
  - «cafés et bistrots» («Forum des Nouveaux», «Le salon de médiation», «Legifer» ...)
- Journaux/Chat d'activité («Bulletin des patrouilleurs»)

Wikipedia As diversified Corpus

Cafés, bistros et espaces généraux

- Le **Forum des nouveaux**, pour poser toutes vos questions à propos de Wikipédia ou son fonctionnement, et y obtenir l'aide de bénévoles.
- Le **Bistro** est un endroit pour se détendre, pour discuter du projet Wikipédia et se renseigner.
- Le **Bistro des non-francophones** permet de discuter à propos de la Wikipédia en français quand on ne parle pas, ou mal, le français. Il permet également, au besoin, de poser des questions sur les pays francophones et leur culture.
- L'**Oracle** est là pour essayer de répondre à votre curiosité sur des *questions d'ordre encyclopédique*.
- **Legifer**. Pour discuter et poser des questions sur les éventuels problèmes légaux posés par le contenu de Wikipédia.
- Les **Questions techniques**, c'est l'endroit où se rendre pour trouver des réponses à ses difficultés ou interrogations lors de sa participation à Wikipédia.
- Le **Salon de médiation**, c'est l'endroit où se rendre pour résoudre dans un cadre serein des conflits, par le biais de la médiation.
- La **WikiPermanence**, c'est une rencontre physique avec des wikipédiens pour ceux qui souhaitent acquérir des connaissances et des conseils sur le fonctionnement de Wikipédia.

Modifier

Bulletins

- Le **bulletin des administrateurs** permet aux admins et aux autres contributeurs de discuter de leurs actions, de leurs frustrations et de leurs explications.
- Le **bulletin des bureaucrates** permet aux bureaucrates de coordonner leurs actions.
- La **Coordination des arbitres**, pour les échanges au sein du comité d'arbitrage.
- Le **bulletin des patrouilleurs** permet aux patrouilleurs et aux autres contributeurs de discuter de leurs frustrations et de leurs explications ainsi que de coordonner leurs actions.

Modifier

Machine à café des Ateliers

Écriture [modifier le code]

- L'**Atelier de lecture** est l'endroit pour faire relire les
label.
- L'**Atelier de relecture** est l'endroit pour faire relire le
- L'**Atelier du français** traite du français et de ses varia
- Le **Salon des grammairiens** répondra à toutes vos
français.
- L'**Atelier typographique** accueille les questions d'or
relative à la typographie.
- L'**Atelier de toponymie** accueille les questions sur le

Image [modifier le code]

- L'**Atelier graphique** a pour objectif d'améliorer et de
Wikipédia.
- L'**Atelier photographique**, pour la recherche et le tr
- L'**Atelier identification** vous aide quand vous avez u

Technique [modifier le code]

- L'**Atelier TeX** a pour objectif d'aider et d'améliorer l
- L'**Atelier accessibilité** a pour but d'informer les contr
d'adapter le contenu de Wikipédia, en particulier pour les

Recherche [modifier le code]

- Le **Wikilab** vise à coordonner les projets de recherche

Les discussions Wikipedia

WIKIPÉDIA
L'encyclopédie libre

Accueil
Portails thématiques
Article au hasard
Contact

Contribuer
Débuter sur Wikipédia
Aide
Communauté
Modifications récentes
Faire un don

Imprimer / exporter
Créer un livre
Télécharger comme PDF
Version imprimable

Outils
Pages liées
Suivi des pages liées
Importer un fichier
Pages spéciales
Adresse de cette version
Information sur la page
Élément Wikidata
Citer cette page

Autres langues
Afrikaans
العربية

Article | Discussion | Lire | Modifier | Modifier le code

Traitement automatique du langage naturel

Le **traitement automatique du langage naturel** ou **de la langue naturelle** (abr. *TALN*) ou **des langues** (abr. *TAL*) est une discipline à la frontière de l'intelligence artificielle, qui concerne l'application de programmes et techniques informatiques à tous les aspects du langage humain[1]. Ainsi, le TAL **linguistique**.

Sommaire [masquer]

1 Histoire
2 TAL statistique
3 Les applications TAL
4 Voir aussi
4.1 Articles connexes
4.2 Liens externes
4.3 Bibliographie
4.4 Références

Histoire [modifier | modifier le code]

L'histoire du TAL commence dans les années 1950, bien que l'on puisse trouver des travaux antérieurs. En 1950, Alan Turing éditait un article célèbre s intelligence » qui propose ce qu'on appelle à présent le test de Turing comme critère d'intelligence. Ce critère dépend de la capacité d'un programme dans une conversation écrite en temps réel, de façon suffisamment convaincante que l'interlocuteur humain ne peut distinguer sûrement — sur la bas s'il interagit avec un programme ou avec un autre vrai humain.
L'expérience de Georgetown en 1954 comportait la traduction complètement automatique de plus de soixante phrases russes en anglais. Les auteurs ou cinq ans, la traduction automatique ne serait plus un problème[2].

Pendant les années 1960, SHRDLU, un système de langage naturel appelé « blocks world » dont la base était des vocabulaires relativement restreints, les chercheurs à l'optimisme.
Cependant, le progrès réel était beaucoup plus lent, et après le rapport ALPAC (en) de 1966, qui constatait qu'en dix ans de recherches les buts n'avaie considérablement réduite.

ELIZA était une simulation à la manière de la psychothérapie rogérienne, écrite par Joseph Weizenbaum entre 1964 à 1966. N'employant presque aucu l'émotion humaine, ELIZA parvenait parfois à offrir un semblant stupéfiant d'interaction humaine. Quand le « patient » dépassait la base de connaissan

Les discussions Wikipedia

WIKIPÉDIA
L'encyclopédie libre

Accueil
Portails thématiques
Article au hasard
Contact

Contribuer
Débuter sur Wikipédia
Aide
Communauté
Modifications récentes
Faire un don

Imprimer / exporter
Créer un livre
Télécharger comme PDF
Version imprimable

Outils
Pages liées
Suivi des pages liées
Importer un fichier
Pages spéciales
Adresse de cette version
Information sur la page

Langues

Article | Discussion | Lire | Modifier le code | Ajouter un sujet

Discussion:Traitement automatique du langage naturel

Autres discussions [liste]
Suppression - Neutralité - Droit d'auteur - Article de qualité - Bon article - Lumière sur - À faire - Archives

Cet article est indexé par les projets Informatique, Langues.

Les projets ont pour but d'enrichir le contenu de Wikipédia en aidant à la coordination du travail des contributeurs. Vous pouvez mo article ou visiter les pages de projets pour prendre conseil ou consulter la liste des tâches et des objectifs.

Évaluation de l'article « **Traitement automatique du langage naturel** »

Cet article comporte une liste de tâches suggérées :

Fusion abandonnée entre Linguistique informatique et Traitement automatique du langage naturel [modifier le code]

Discussion transférée depuis Wikipédia:Pages à fusionner

Si on en croit les résumés introductifs, c'est la même chose. --Rinaku (d · c) 11 janvier 2013 à 23:58 (CET)

⊕ **Pour**. Je confirme, c'est la même chose. --Pierre Rudloff (d) 12 janvier 2013 à 03:42 (CET)

⊕ **Pour** Même chose. Je serais favorable à une fusion sous le titre **traitement automatique du langage naturel** qui est, d'après mon expérience dans le monde académique. --Ænthaüs (d) 18 janvier 2013 à 11:16 (CET)

> +1, je ne connaissais que cette seconde formulation. --Rinaku (d · c) 19 janvier 2013 à 15:01 (CET)

⊕ **Pour**, d'autant que l'association entre Linguistique informatique et Langage informatique me parait artificielle. Bmathis (d) 2 février 2013 à 16:1

⊖ **Contre** La linguistique computationnelle (ou informatique) est la discipline scientifique qui étudie le phénomène linguistique (grammaire, séman informatiques (règles chomskyennes, grammaires formelles, etc). Le Traitement Automatique du Langage Naturel est la discipline scientifique qui u automatique au sens large (apprentissage automatique, classification, traitement du signal) pour réaliser des traitements sur le langage naturel (so cela concerne par exemple la transcription (reconnaissance de la parole), la classification (classification de document textuels), l'étiquetage. Certai automatique (TA) peuvent être hybrides et utiliser les deux méthodes. Voir pour mieux comprendre la distinction Natural Language Processing (eq T sur Wikipedia en. Bublegun (d) 2 février 2013 à 17:30 (CET)

> Si c'est vrai, alors le titre de l'article Linguistique informatique ne correspond pas à son contenu qui, lui, traite bien du TAL. Il faudrait donc fusion

Les discussions Wikipedia

Accessibilité et quantité des données

- "Forum de discussion" libre distribué publiquement (Creative Commons by-sa)
- existe dans presque toutes les langues $\rightarrow$ objet d'étude international

Richesse des métadonnées

- thématique (portail thématique, article associé)
- accès aux connaissances partagées (article associé)
- degré de subjectivité (appel au calme, etc.)
- informations sur le locuteur (statut dans la communauté, participation à la Wikipedia, possibilité de profilage)

Procédure de constitution

1. Extraction des discussions depuis le dump
   sauvegarde globale des pages courantes de la Wikipedia française (archive frwiki-20150512-pages-meta-current#.xml.bz2 diffusée librement sur la page http ://dumps.wikimedia.org/frwiki/20150512/)
2. Sélection des discussions "à garder"
3. Analyse des objets textuels constitutifs de chaque discussion : sections, messages
4. Conversion selon la TEI-P5
5. Analyse syntaxique automatique (Talismane, Urieli 2013)

Procédure de constitution - sélection des discussions

/<title>Discussion/ sur le dump du 20150512 : 3 487 480

Discussions portant sur un utilisateur	1 990 927	57%
Discussions portant sur un article	**1 496 553**	**43%**
Discussions redirigées vers une autre discussion	116 432	8%
Discussions vides ou contenant moins de 2 mots	1 013 791	68%
Discussions retenues	**366 326**	**24%**

Procédure de constitution - structuration des discussions

Des discussions à la norme TEI-P5

1. Extraction des méta-données
2. Structuration en sections (fils) et messages (posts)
3. Délimitation des différentes contributions : 1 message - 1 date de publication
4. Évaluation de l'extraction

Procédure de constitution - structuration des discussions

Des discussions à la norme TEI-P5

1. **Extraction des méta-données**
2. Structuration en sections (fils) et messages (posts)
3. Délimitation des différentes contributions : 1 message - 1 date de publication
4. Évaluation de l'extraction

Procédure de constitution - structuration des discussions

Extraction des méta-données

Procédure de constitution - structuration des discussions

Des discussions à la norme TEI-P5

1. Extraction des méta-données
2. **Structuration en sections (fils) et messages (posts)**
3. **Délimitation des différentes contributions : 1 message - 1 date de publication**
4. Évaluation de l'extraction

Procédure de constitution - structuration des discussions

Structuration des discussions


Fusion abandonnée entre Linguistique informatique et Traitement automatiqu

Discussion transférée depuis Wikipédia:Pages à fusionner

Si on en croit les résumés introductifs, c'est la même chose. --Rinaku (d · c) 11 janvier 2013 à

⊕ **Pour**. Je confirme, c'est la même chose. --Pierre Rudloff (d) 12 janvier 2013 à 03:42 (C

⊕ **Pour** Même chose. Je serais favorable à une fusion sous le titre **traitement automat**
dans le monde académique. --Ænthaüs (d) 18 janvier 2013 à 11:16 (CET)

+1, je ne connaissais que cette seconde formulation. --Rinaku (d · c) 19 janvier 2013 à

⊕ **Pour**, d'autant que l'association entre Linguistique informatique et Langage informati

⊖ **Contre** La linguistique computationnelle (ou informatique) est la discipline scientifiqu
informatiques (règles chomskyennes, grammaires formelles, etc). Le Traitement Automa
automatique au sens large (apprentissage automatique, classification, traitement du sigr
cela concerne par exemple la transcription (reconnaissance de la parole), la classificatio
automatique (TA) peuvent être hybrides et utiliser les deux méthodes. Voir pour mieux cc
sur Wikipedia en. Bublegun (d) 2 février 2013 à 17:30 (CET)

Si c'est vrai, alors le titre de l'article Linguistique informatique ne correspond pas à so
informatique dans TAL mais laisser la page Linguistique informatique à l'état d'ébauc

Je suis assez d'accord avec cela. Le contenu de Linguistique informatique est effec
(CET)

⊖ ~~**Neutre** La discussion a déjà eu lieue sur wikipedia en. Pour beaucoup de spécialist~~
~~Comme indiqué plus haut, le CL a une coloration plus linguistique, mais je doute que des~~
~~académique, on trouve par exemple la conférence CICLing qui traite explicitement des~~
~~contre, il faudra sûrement modifier le contenu pour bien expliciter les liens etre éventuel~~

Les conférences majeures sont effectivement souvent indifférenciées (telles Cicling
CoNLL qui est orientée traitement automatique, ou INLG qui est elle très orientée Com

```
<div id="1" level="1">
  <head>Fusion abandonnée entre [[Linguistique informati
  <sp id="1" who="Rinaku" when="11-01-2013-23:58" intera
    <p> Discussion transférée depuis ) 11 janvier 2013 à
  </sp>
  <sp id="2" who="Rudloff" when="12-01-2013-03:42" inter
    <p> pour. Je confirme, c'est la même chose. --Pierre
  </sp>
  <sp id="3" who="Ænthaüs" when="18-01-2013-11:16" inter
    <p> pour Même chose. Je serais favorable à une fusio
    formulation la plus usitée, du moins dans le monde a
  </sp>
  <sp id="4" who="Rinaku" when="19-01-2013-15:01" intera
    <p> +1, je ne connaissais que cette seconde formulat
  </sp>
  <sp id="5" who="Bmathis" when="02-02-2013-16:15" inter
    <p> pour, d'autant que l'association entre Linguisti
    (CET).</p>
  </sp>
  <sp id="6" who="Bublegun" when="02-02-2013-17:30" inte
    <p> contre La linguistique computationnelle (ou info
    structure) avec des outils informatiques (règles cho
    scientifique qui utilise des méthodes de traitement
    réaliser des traitements sur le langage naturel (sou
    parole), la classification (classification de docume
    hybrides et utiliser les deux méthodes. Voir pour mi
    LC) sur Wikipedia en. Bublegun  2 février 2013 à 17:
  </sp>
  <sp id="7" who="Rinaku" when="03-02-2013-12:06" intera
    <p> Si c'est vrai, alors le titre de l'article ) 3 f
  </sp>
  <sp id="8" who="Bublegun" when="04-02-2013-21:04" inte
    <p>  Je suis assez d'accord avec cela. Le contenu de
    Bublegun  4 février 2013 à 21:04 (CET)</p>
  </sp>
  <sp id="9" who="Xiawi" when="03-02-2013-22:12" interac
    <p> neutre La discussion http://en.wikipedia.org/wik
```

Procédure de constitution - structuration des discussions

Des discussions à la norme TEI-P5

1. Extraction des méta-données
2. Structuration en sections (fils) et messages (posts)
3. Délimitation des différentes contributions : 1 message - 1 date de publication
4. **Évaluation de l'extraction**

Évaluation de la constitution

Évaluation de l'extraction

- 7 discussions évaluées manuellement : 413 messages et 47284 mots
- précision = 0,92, rappel = 0,95

Bruit 3 messages vides ; 5 messages scindés en 2 ; 25 messages fusionnant 2 ou 3 messages

Silence 23 messages absents

Évaluation de la constitution - exemples d'erreur

sens du mot Lehi

66	0	Le truc c'est qu'en hébreu, et en arabe, la frontière est parfois mince entre un acronyme et une abréviation. "Fatah" ou "Hamas" par exemple sont-ils acronymes ou abréviations ? --Markov 5 septembre 2006 à 02:04 (CEST)
67	1	Je pense que ce sont res rétro-acronymes : on prend un mot qui veut dire quelque chose, et on invente un sigle dont l'acronyme deviendra le mot choisit. C'est ça ? Par contre je ne crois pas que Lehi veuille dire quelque chose en Hébreu. Tu comprend l'hébreu, Markov ? Christophe Cagé - liste de mes articles 5 septembre 2006 à 07:06 (CEST)
68	3	Euh, "ksat, ksat", (très peu), notions de base. --Markov 8 septembre 2006 à 11:02 (CEST)
69	2	A ma connaissance Lehi, ne veut rien dire mais je ne suis pas la référence. En cherchant sur google j'ai trouvé que c'était un lieu dit où les Philistins et les hébreux s'affrontèrent mais cela ne prouve pas la volonté de faire le lien... Ceedjee 5 septembre 2006 à 07:45 (CEST)
70	0	Ben si les hébreux ont gagnés, c'est en tout cas un indice. C'est le cas ? Christophe cagé Au delà du renommage, il serait utile de pourvoir faire la distinction entre une abréviation par acronymie ou par un sigle. Personnellement, je n'ai jamais vu d'abréviations écrits en lettres capitales, c'est de cette façon que j'avais pensé que LEHI était un sigle. Maintenant le cas de l'hébreux semblent particulier, je ne suis pas apte à trancher (surtout pour une Rétro-acronymie). VIGNERON * 5 septembre 2006 à 08:56 (CEST)
71	1	Tu a raison. J'ai mis LEHI en majuscule, parceque c'est la graphie de Schattner, mais d'autres historiens mettent des minuscules, je crois. Il faut que je vérifie. Sinon, j'ai demandé son avis à Franckiz. Il a un niveau de base en hébreu (mais ce n'est pas un expert, sauf erreur) Christophe Cagé Non, LEHI ne veut rien dire en hébreu. En tout cas rien qui puisse avoir un rapport avec le contexte. ("Léhi" : "Va" à l'impératif féminin). zeeev 5 septembre 2006 à 15:29 (CEST)
72	1	De plus le rajout du "e" dans l'abréviation vient du fait que l'importance des voyelles en hébreu est secondaire. Ce sont les consomnes seulement qui composent la racine du mot hébreu. On appelle donc cette organisation "Léhi", mais il est vrai qu'en toute logique, on aurait pu l'appeler "Lahi" ou "Lohi".....zeeev 5 septembre 2006 à 15:37 (CEST)
73	2	Ta réponse n'est que partielle. Tu dit "pourquoi pas un e". Certe, mais pourquoi pas LHI ? Ca ne se fait pas, en hébreu ?Christophe Cagé - liste de mes articles 6 septembre 2006 à 06:14 (CEST)

sens du mot Lehi [modifier le code]

Le truc c'est qu'en hébreu, et en arabe, la frontière est parfois mince entre un acronyme et une abréviation. "Fatah" ou "Hamas" par exemple sont-ils acronymes ou abréviations ? --Markov (discut.) 5 septembre 2006 à 02:04 (CEST)

Je pense que ce sont res rétro-acronymes : on prend un mot qui veut dire quelque chose, et on invente un sigle dont l'acronyme deviendra le mot choisit. C'est ça ? Par contre je ne crois pas que Lehi veuille dire quelque chose en Hébreu. Tu comprend l'hébreu, Markov ? Christophe Cagé - liste de mes articles 5 septembre 2006 à 07:06 (CEST)

Euh, *ksat, ksat*, (très peu), notions de base. --Markov (discut.) 8 septembre 2006 à 11:02 (CEST)

A ma connaissance Lehi, ne veut rien dire mais je ne suis pas la référence. En cherchant sur google j'ai trouvé que c'était un lieu dit où les Philistins et les hébreux s'affrontèrent mais cela ne prouve pas la volonté de faire le lien... Ceedjee *contact* 5 septembre 2006 à 07:45 (CEST)

Ben si les hébreux ont gagnés, c'est en tout cas un indice. C'est le cas ? Christophe cagé

Au delà du renommage, il serait utile de pourvoir faire la distinction entre une abréviation par acronymie ou par un sigle. Personnellement, je n'ai jamais vu d'abréviations écrits en lettres capitales, c'est de cette façon que j'avais pensé que LEHI était un sigle. Maintenant le cas de l'hébreux semblent particulier, je ne suis pas apte à trancher (surtout pour une Rétro-acronymie). VIGNERON * discut. 5 septembre 2006 à 08:56 (CEST)

Tu a raison. J'ai mis LEHI en majuscule, parceque c'est la graphie de Schattner, mais d'autres historiens mettent des minuscules, je crois. Il faut que je vérifie. Sinon, j'ai demandé son avis à Franckiz. Il a un niveau de base en hébreu (mais ce n'est pas un expert, sauf erreur) Christophe Cagé

Non, LEHI ne veut rien dire en hébreu. En tout cas rien qui puisse avoir un rapport avec le contexte. ("Léhi" : "Va" à l'impératif féminin). zeeev 5 septembre 2006 à 15:29 (CEST)

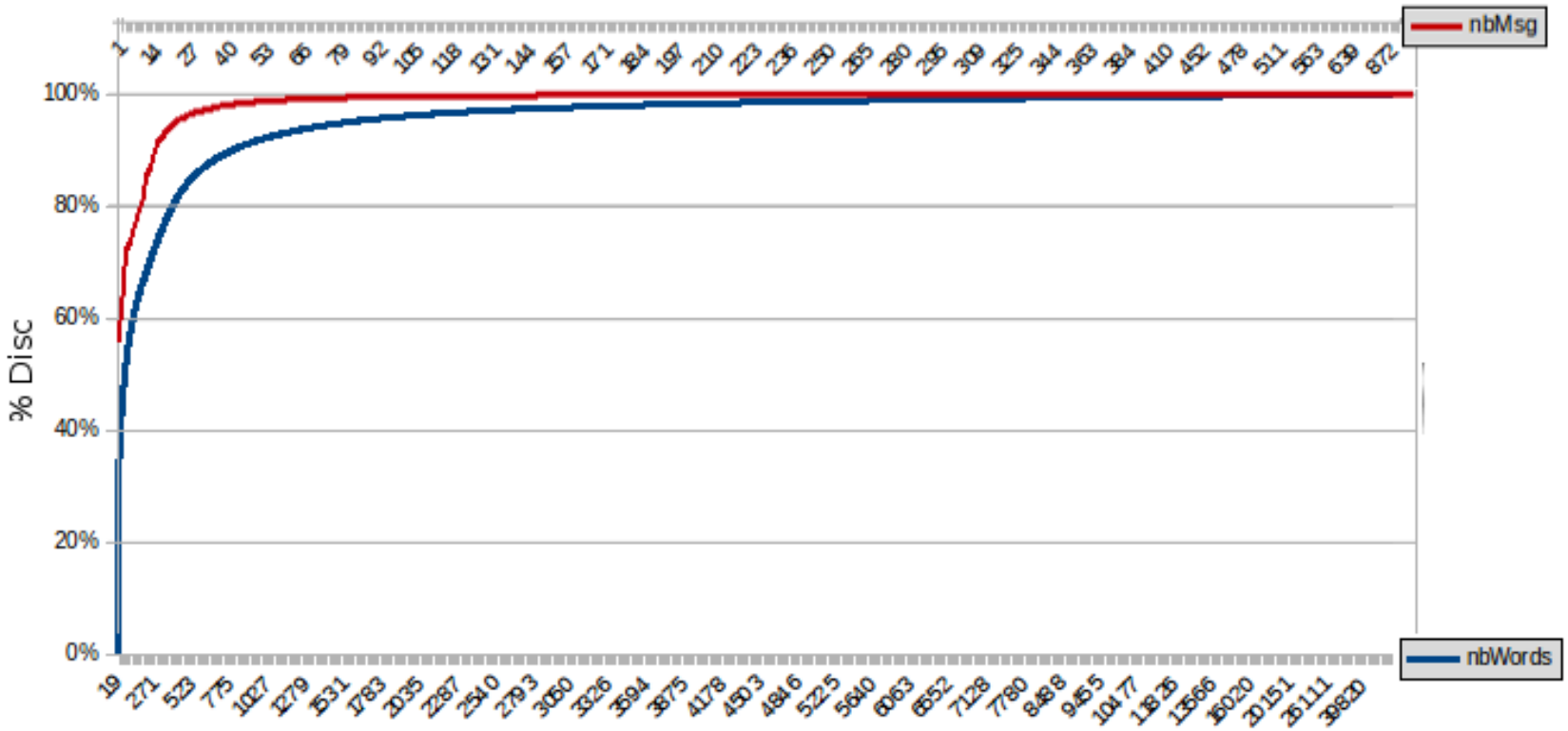
De plus le rajout du "e" dans l'abréviation vient du fait que l'importance des voyelles en hébreu est secondaire. Ce sont les consomnes seulement qui composent la racine du mot hébreu. On appelle donc cette organisation "Léhi", mais il est vrai qu'en toute logique, on aurait pu l'appeler "Lahi" ou "Lohi".....zeeev 5 septembre 2006 à 15:37 (CEST)

Ta réponse n'est que partielle. Tu dit "pourquoi pas un e". Certe, mais pourquoi pas LHI ? Ca ne se

Caractéristiques globales - longueur

discussions	sections	messages	mots
366 326	1 024 351	3 022 240	159 578 279

202 856 (55%) discussions ne contenant qu'un message
181 503 (50%) contenant moins de 53 mots

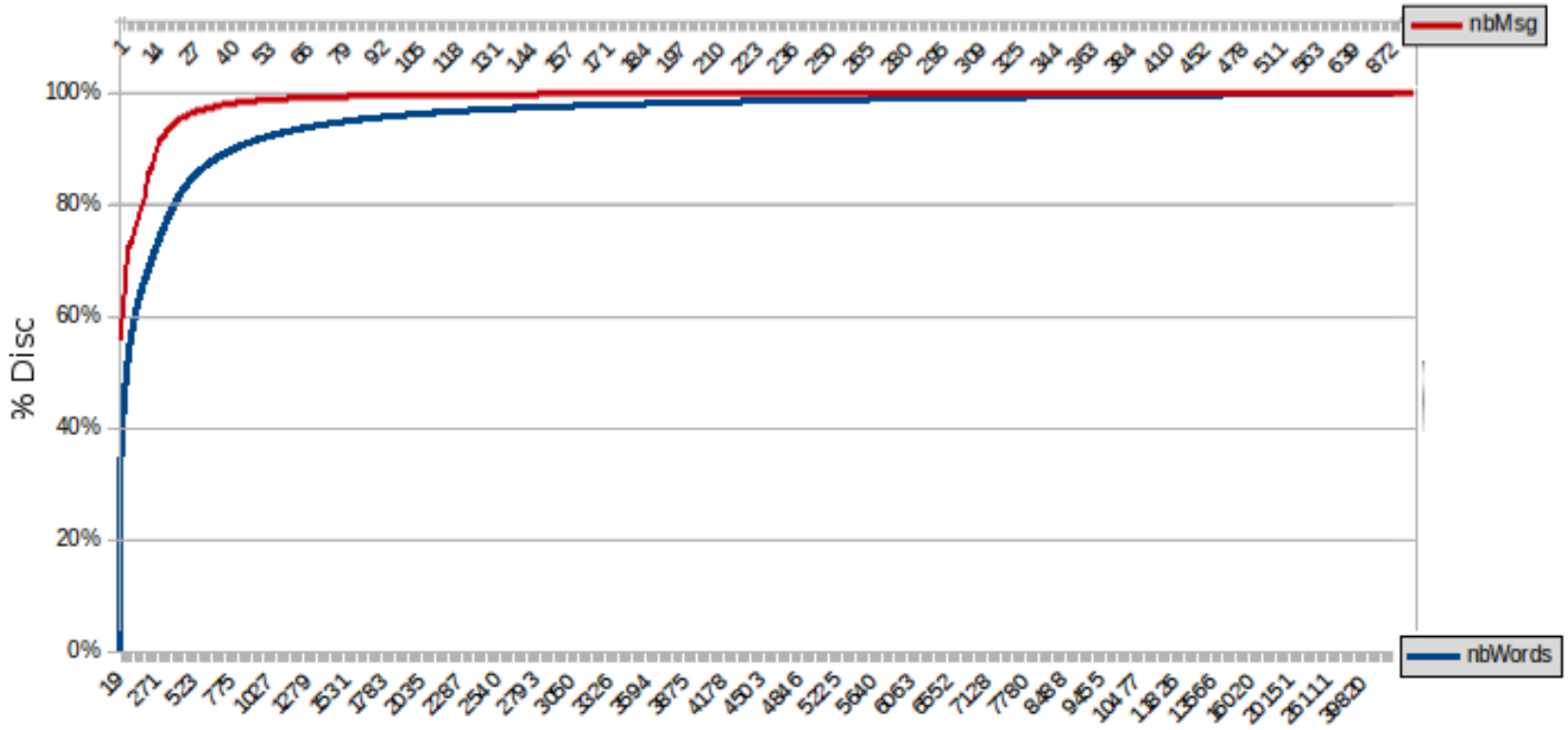


Exemple de discussion mono message

Caractéristiques globales - longueur

Des discussions allant jusqu'à 1 143 messages et 148 968 mots

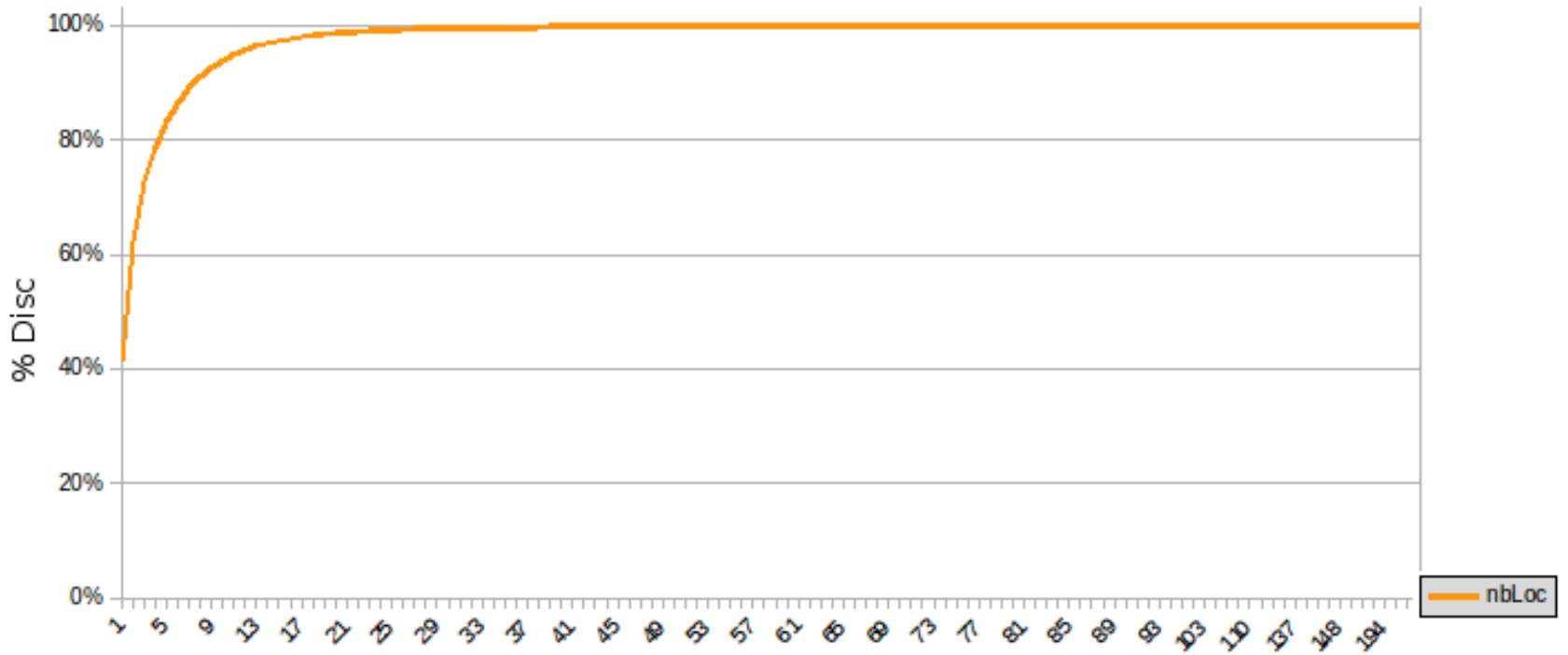
"*Opposition au mariage homosexuel en France part1*"



Caractéristiques globales - interactions

150 603 (41%) monologues
40 413 (10%) impliquant entre 8 et 228 locuteurs différents*

"*Opposition au mariage homosexuel en France part1*" : 93 locuteurs (anonyme = 1)



80% de contributions anonymes

* Discussion sur l'admissibilité de la page "Mickaël Vendetta"

Plan

1 Motivations : Web As Corpus

2 Wikipedia As Corpus

3 Premières analyses
- Méthodologie générale
- Analyses hypothesis-driven
- Analyses data-driven

4 Conclusions et perspectives

Étude contrastive de plusieurs genres pour **caractériser** le genre "discussion"

Combinaison d'approches

- Approche hypothesis-driven : permet d'examiner les caractéristiques typiques à certains textes dans de nouveaux textes
- Approche data-driven : permet de découvrir les caractéristiques typiques des corpus sans connaissances à priori

Approches hypothesis-driven et data-driven

Approche *hypothesis-driven* : mesurer des caractéristiques *a priori* spécifiques

- Extraction de patrons → degré de "déviance", formules d'ouverture
- Projection de lexique → traces de subjectivité

Approche data-driven : découvrir les différences entre les textes du corpus

Tâche de classification automatique supervisé : le logiciel apprend à identifier les classes de textes (càd les sous-corpus) et en donne les traits les plus typiques (caractéristiques)

- Importance du choix des traits utilisés. **Proposition** : N-grams lexicaux, morphologiques, syntaxiques

Corpus de comparaison

Constitués

Corpus écrits	No tokens	Description
Rue89	2 192 995	Presse en ligne
AgoraVox	4 099 662	Media citoyen
Forum Santé	236 368 151	Forum de discussion
		2 585 188 messages
WikiDiscussion (2015)	132 406 816	Forum de discussion
		3 022 240 messages
WikiArticles (2013)	226 207 672	Articles encyclopédiques

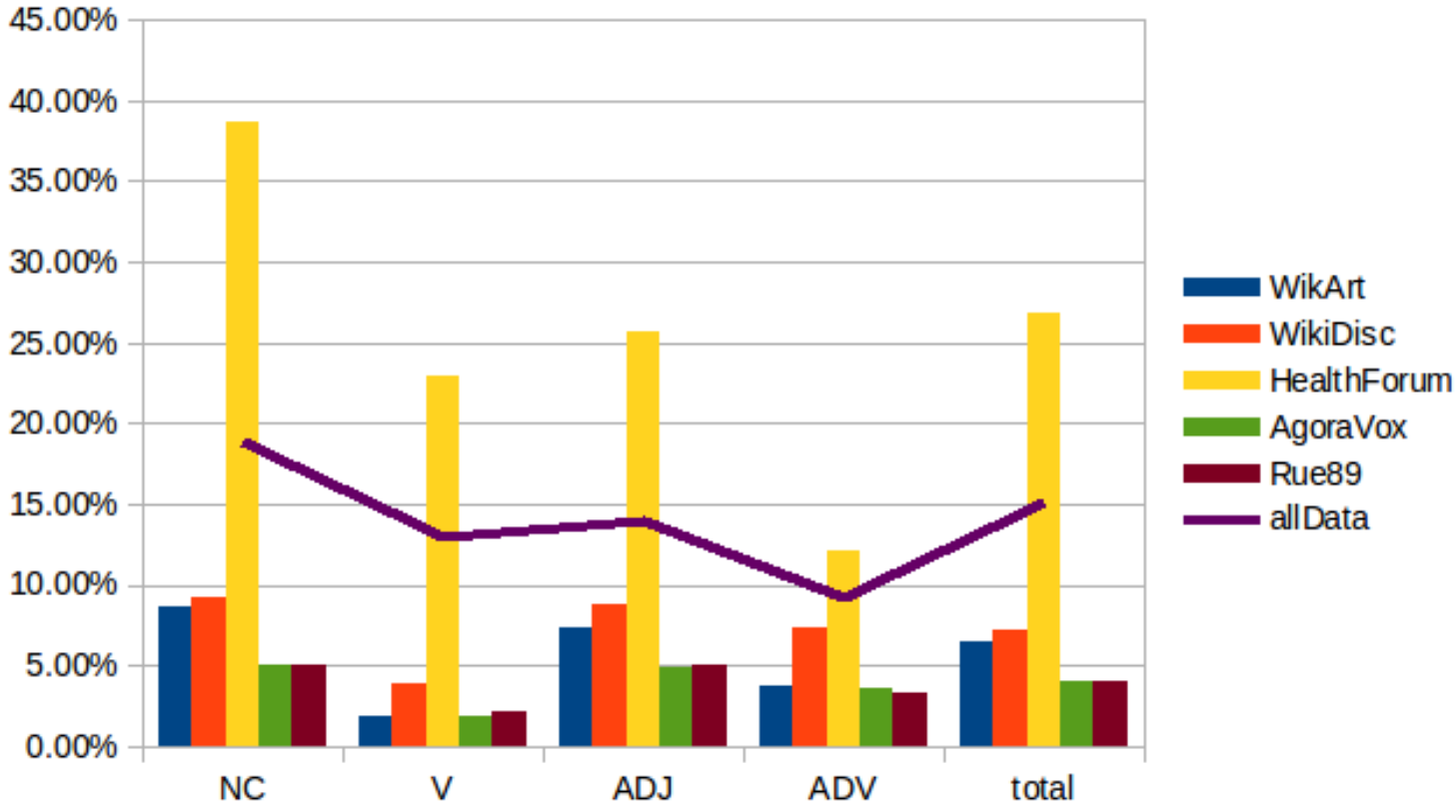
Parsés

Tous les corpus ont été analysés syntaxiquement avec Talismane (version 1.8.5b, beam de 5, modèle svn)

Évaluation du degré de "déviance" de l'écriture

Méthode : mesurer le taux de mots inconnus

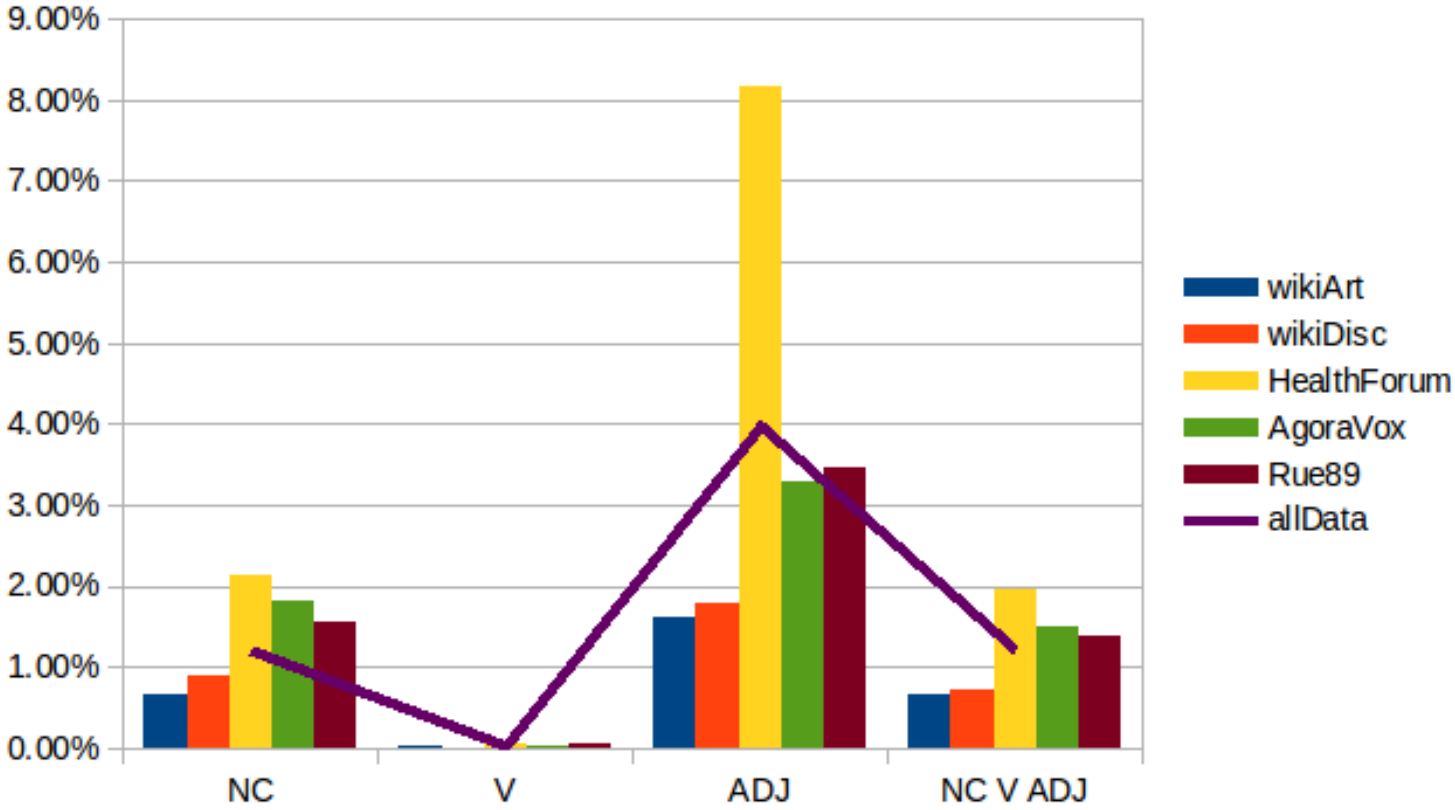
% de mots inconnus, càd sans lemme associé selon Talismane (Lefff) dans les catégories nom, verbe, adjectif et adverbe



Traces de subjectivité

Méthode : projection d'un lexique des affects (Augustyn et al. 2008)

% de Noms, Verbes et Adjectifs mentionnés dans le lexique des affect



Formules d'ouverture : WikiDiscussions vs. Forum Santé

Méthode : Extraction des n-grams (n < 4) de début de message

WikiDisc	Init msg/phr	‰ Msg
Discussions	99.9	18
Avis	99.7	17
supprimer	97.1	6
conserver	95.7	5
Neutre	98.9	4
Votes	99.8	3
Signalé_par	98.8	3
des_articles_admissibles	99.2	3
Si_vous_êtes	15.6	2
Il_me_semble	32.8	2
Bilan	97.9	2
Il_y_a	23.4	2
Merci	28.2	2
Ce_n'_est	22.8	2
pourBA	98.7	2
Bon	53.6	2
Je_viens_de	60.8	2
Je_ne_vois	32.3	1
Je_suis_d'accord	59.7	1
En_effet	23.0	1
Je_pense_qu'	30.1	1
Effectivement	57.4	1
pour	88.1	1
Je_ne_sais	29.9	1

Formules d'ouverture : WikiDiscussions vs. Forum Santé

Méthode : Extraction des n-grams (n < 4) de début de message

WikiDisc	Init msg/phr	%o Msg
Discussions	**99.9**	18
Avis	**99.7**	17
supprimer	**97.1**	6
conserver	**95.7**	5
Neutre	**98.9**	4
Votes	**99.8**	3
Signalé_par	**98.8**	3
des_articles_admissibles	**99.2**	3
Si_vous_êtes	15.6	2
Il_me_semble	32.8	2
Bilan	**97.9**	2
Il_y_a	23.4	2
Merci	28.2	2
Ce_n'_est	22.8	2
pourBA	**98.7**	2
Bon	53.6	2
Je_viens_de	60.8	2
Je_ne_vois	32.3	1
Je_suis_d'accord	59.7	1
En_effet	23.0	1
Je_pense_qu'	30.1	1
Effectivement	57.4	1
pour	88.1	1
Je_ne_sais	29.9	1

- **"code interne"**
- recherche d'accord

Formules d'ouverture : WikiDiscussions vs. Forum Santé

Méthode : Extraction des n-grams ($n < 4$) de début de message

forum Santé	Init msg/phr	%o Msg
Coucou_les_filles	99.0%	25
coucou_les_filles	99.3%	14
Coucou	89.0%	8
Salut_les_filles	99.2%	7
Bonjour	87.9%	7
Bonjour_les_filles	99.0%	7
coucou	92.0%	5
salut_les_filles	99.0%	4
bonjour_les_filles	99.3%	3
Coucou_les	99.0%	3
Merci	47.4%	2
Merci_les_filles	80.0%	2
bonjour	90.9%	2
Salut	89.2%	2
Bonsoir_les_filles	99.3%	2
coucou_les	99.5%	2
Oui	51.4%	2
Oui_c'_est	57.2%	2
merci_les_filles	85.3%	2
Comment_allez_vous	44.0%	2
merci	57.6%	2
oui_c'_est	66.7%	1

Formules d'ouverture : WikiDiscussions vs. Forum Santé

Méthode : Extraction des n-grams (n < 4) de début de message

forum Santé	Init msg/phr	%o Msg
Coucou_les_filles	**99.0%**	25
coucou_les_filles	**99.3%**	14
Coucou	**89.0%**	8
Salut_les_filles	**99.2%**	7
Bonjour	**87.9%**	7
Bonjour_les_filles	**99.0%**	7
coucou	**92.0%**	5
salut_les_filles	**99.0%**	4
bonjour_les_filles	**99.3%**	3
Coucou_les	**99.0%**	3
Merci	47.4%	2
Merci_les_filles	80.0%	2
bonjour	**90.9%**	2
Salut	**89.2%**	2
Bonsoir_les_filles	**99.3%**	2
coucou_les	**99.5%**	2
Oui	51.4%	2
Oui_c'_est	57.2%	2
merci_les_filles	**85.3%**	2
Comment_allez_vous	44.0%	2
merci	57.6%	2
oui_c'_est	66.7%	1

- Spécificité des locuteurs
- Conversation

Premières conclusions

Première caractérisation des WikiDiscussions

- Des discussions "bien écrites"
- A priori peu empruntes d'*affects*
- A la recherche d'un consensus

Approche data-driven

Découvrir les différences entre les textes du corpus

Tâche de classification automatique supervisé : le logiciel apprend à identifier les classes de textes (càd les sous-corpus) et en donne les traits les plus typiques (caractéristiques)

- Importance du choix des traits utilisés.
- Proposition : N-grams lexicaux, morphologiques, syntaxiques

Approche data-driven

Méthode classique

Calcul des différences entre corpus à l'aide de *mots-clés* (Scott & Tribble 2006)

- Mots statistiquement spécifiques à un corpus par rapport à un corpus de référence
- Informatifs sur la thématique et style du corpus
- Les (n-grams) de mots également efficaces pour la classification des textes du même domaine (Laippala et al. 2015a)

Inconvénients

- Les mots ne reflètent que difficilement la variation morphologique et syntaxique
- Inefficace pour une classification non thématique ou inter-domaines

Caractériser les corpus sans se limiter à un domaine particulier

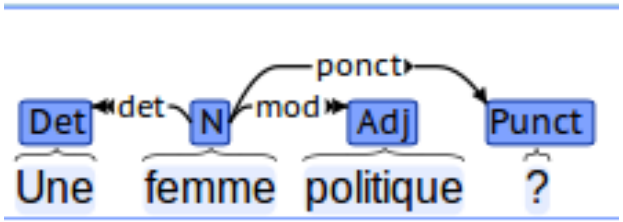
Caractériser le web nécessite à caractériser les corpus sans se limiter à un domaine particulier

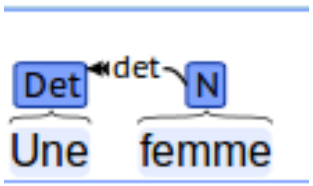
- Une première question : Est-ce possible ?
- Y a-t-il des similarités entre les textes similaires que l'information lexicale ne couvre pas ?
  - Articles journalistiques sur des thèmes différents ?
  - Discussions en ligne sur des thèmes différents ?
- Y a-t-il des caractéristiques morphologiques et syntaxiques partagées par des textes appartenant au même genre ?

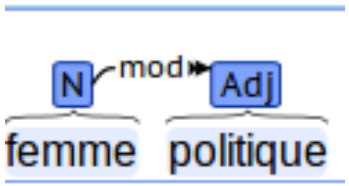
→ Notre solution : N-grams syntaxiques

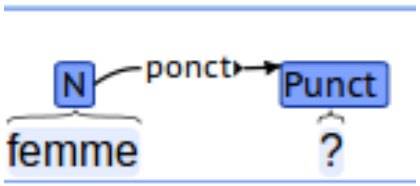
N-grams syntaxiques

- Petits sous-arbres d'une analyse syntaxique en dépendance
- Suivent les relations de dépendance ⇒ pas nécessairement linéaires
- Méthode appliquée à l'origine à l'anglais (Goldberg & Orwant 2013)
- Adaptation au finnois pour le Finnish Internet Parsebank + publication du pipeline (Kanerva & al. 2014)
- Adaptation au français (Laippala & Ho-Dac 2015)



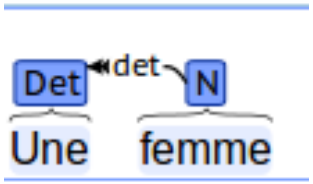


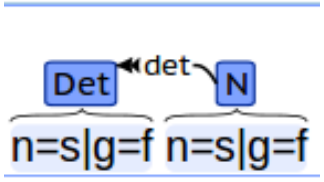


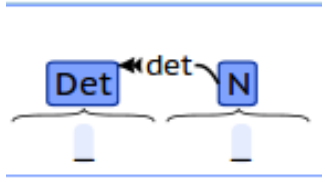


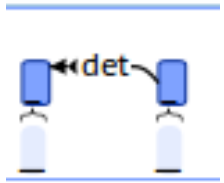
N-grams syntaxiques : paramétrage des niveaux de granularité et de lexicalisation

Conservation des lexèmes, traits morphologiques, classes morphologiques, relations de dépendances







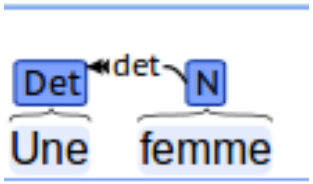


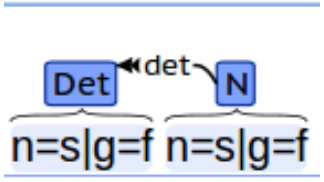
Analyse détaillée
Qté d'informations ++

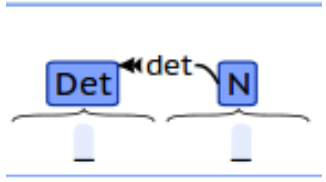
[-] Analyse abstraite
Qté d'information - -

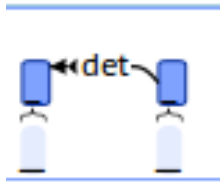
N-grams syntaxiques : paramétrage des niveaux de granularité et de lexicalisation

Conservation des lexèmes, traits morphologiques, classes morphologiques, relations de dépendances









Analyse détaillée
Qté d'informations ++

[-] Analyse abstraite
Qté d'information - -

Pour nous : Méthode d'analyse au-delà du niveau lexical pour faire émerger (classer selon) des structures spécifiques (et non des termes ou des collocations spécifiques)

Travaux antérieurs sur les N-grams syntaxiques en Finnois

Détection des traductions faites par humains (HT) et des traductions automatiques (MT) (Laippala & al. 2015a)

Deux tâches

HT Texte créé en finnois vs. texte traduit en finnois par un humain

MT Texte créé en finnois vs. texte traduit en finnois par une machine

Travaux antérieurs sur les N-grams syntaxiques en Finnois

Détection des traductions faites par humains (HT) et des traductions automatiques (MT) (Laippala & al. 2015a)

Deux tâches

HT Texte créé en finnois vs. texte traduit en finnois par un humain

MT Texte créé en finnois vs. texte traduit en finnois par une machine

Résultats pour *out-of-domain*

Bag-of-Words	HT $\rightarrow$ F-mesure 59%	MT $\rightarrow$ F-mesure 73%
N-grams syntaxiques	HT $\rightarrow$ F-mesure 65%	MT $\rightarrow$ F-mesure 92%

Travaux antérieurs sur les N-grams syntaxiques en Finnois

Détection des traductions faites par humains (HT) et des traductions automatiques (MT) (Laippala & al. 2015a)

Deux tâches

HT Texte créé en finnois vs. texte traduit en finnois par un humain

MT Texte créé en finnois vs. texte traduit en finnois par une machine

Résultats pour *out-of-domain*

Bag-of-Words	HT $\rightarrow$ F-mesure 59%	MT $\rightarrow$ F-mesure 73%
N-grams syntaxiques	HT $\rightarrow$ F-mesure 65%	MT $\rightarrow$ F-mesure 92%

Conclusions

- N-grams syntaxiques semblent mieux s'adapter au domaine

$\Rightarrow$ Similarités entre les textes au-delà du niveau lexical !

Analyse des caractéristiques des corpus à l'aide des N-grams syntaxiques (Laippala & al. 2015b)

- Trois corpus
  - littérature
  - forum de discussion, rubrique sport
  - commentaires d'articles de presse
- Résultats : deux types de caractéristiques mises en avant par les N-grams
  1. Caractéristiques thématiques, telles que noms composés, urls, etc.
     $\Rightarrow$ Similaire aux mots-clés mais à un niveau plus abstrait
  2. Caractéristiques syntaxiques
     - Structures narratives, impératives, négations
     - Ne reflètent pas la thématique, mais...

     $\Rightarrow$ Caractéristiques du discours ! ! (ou registre à la Biber ?)

     :)

Plan d'action pour l'adaptation au français

1. Petit rappel des questions de recherche
2. Corpus utilisés : WikiDiscussion, Forum Santé, Articles Wikipedia
3. Adaptation des N-grams syntaxiques pour le français
4. Classifieur et paramétrage
5. Tâche 1 : distinguer WikiDiscussion $\leftrightarrow$ Forum Santé
   - *Bag-of-words* (termes simples)
   - N-grams de mots
   - N-grams syntaxiques
6. Tâche 2 : distinguer WikiDiscussion $\leftrightarrow$ Articles Wikipedia
   - *Bag-of-lemma* (termes simples)
   - N-grams de mots
   - N-grams syntaxiques
   - Comparaison aux tri-grams lexicaux

Pour vous rappeler nos questions de recherche...

Questions à court terme

- Quelles sont les caractéristiques du genre "discussion" (Wikipedia) ?
- Comment les N-grams syntaxiques fonctionnent pour le français ?

Questions à plus long terme

- Évaluer l'utilité des N-grams syntaxiques dans la description linguistique des corpus
- Identifier des caractéristiques génériques aux genres du web et ainsi définir ces genres

Corpus utilisés

WikiDiscussion ↔ Forum Santé

- WikiDiscussion (12.903.816 tokens, 636.553 phrases)
- Forum Santé (12.182.582 tokens, 1.170.791 phrases)

WikiDiscussion ↔ Articles Wikipedia

- WikiDiscussion (4.634.209 tokens, 232.807 phrases)
- Articles Wikipedia (4.349.085 tokens, 212.826 phrases)

Adaptation des n-grams syntaxiques pour le français

Le processus de génération des N-grams syntaxiques est basé sur le schéma morphosyntaxique de Stanford, présenté en conll09

Modifications nécessaires pour le français

- Les classes morphologiques (jeu de POS Stanford)
- Le format (.tal → .conll09)

Format CONLL09

#	token	lemma	lemma	POS	POS	MORF	MORF	HEAD	HEAD	REL	REL
1	Une	une	une	DET	DET	g=f|n=s	g=f|n=s	2	2	det	det
2	traduction	traduction	traduction	N	N	g=f|n=s	g=f|n=s	0	0	root	root

Format Talismane

#	token	lemma	POS	POS	MORF	HEAD	REL	HEAD	REL
1	Une	une	DET	DET	g=f|n=s	2	det	2	det
2	traduction	traduction	NC	nc	g=f|n=s	0	root	0	root

Classifieur utilisé

Vowpal Wabbit

- Classifieur linéaire développé par Yahoo ! Research puis repris par Microsoft Research (Agarwal & al. 2014)
- `https://github.com/JohnLangford/vowpal_wabbit/wiki`
- Avantages :
  - plus rapide et paramétrable que les classifieurs SVM (*"a fast, scalable, useful learning algorithm."*)
  - donne accès aux traits les plus significatifs pour chaque classe → caractéristtiques linguistiques des corpus à utiliser dans la caractérisation !

Paramétrage du classifieur

Paramétrage de VW

- "Stochastic gradient descent method for training" (ne pas poser de question)
- Division du corpus : 50% entraînement, 50% test
- Taille du segment de texte à classifier : trois phrases (un peu par défaut)
- WikiDiscussion la classe positive

Paramétrage du classifieur

Traduction

- La moitié du corpus est consacrée à l'apprentissage d'un modèle statistique permettant de classer chaque 3 phrases dans la catégorie WikiDiscussion ou pas.
- Les groupes de 3 phrases du reste du corpus sont classifiés selon ce modèle
- Le résultat de cette classification permet d'évaluer le modèle appris

Interprétation des résultats

- F-mesure évaluant l'efficacité du classifieur
- Traits pertinents selon le modèle

bag-of-words, WikiDiscussion vs. Forum Santé

WikiDiscussion, F-mesure 100%

article wikipedia Wikipedia sources seigneur fusion mulot source peuple supprimé habitants références anonyme CET archeos articles bibliographie définition encyclopédie huit matthieu pdd modifs texte panoramix

Forum Santé, F-mesure 99%

enceintes désolée quote poussette lila34 essais nausicaa pma bb douleur 45 zhom ui fiv grossesse ovulation filles écho chéri cher fofo dpo fc gygy récap bisous coucou 74 rigolo emoticone

bag-of-words, WikiDiscussion vs. Forum Santé

Des termes relatifs aux processus de rédaction sous Wikipedia

WikiDiscussion, F-mesure 100%

article wikipedia Wikipedia sources seigneur **fusion** mulot **source** peuple **supprimé** habitants **références anonyme CET** archeos **articles bibliographie définition encyclopédie** huit matthieu pdd **modifs texte** panoramix

Forum Santé, F-mesure 99%

enceintes désolée quote poussette lila34 essais nausicaa pma bb douleur 45 zhom ui fiv grossesse ovulation filles écho chéri cher fofo dpo fc gygy récap bisous coucou 74 rigolo emoticone

bag-of-words, WikiDiscussion vs. Forum Santé

Des termes relatifs aux processus de rédaction sous Wikipedia

WikiDiscussion, F-mesure 100%

article wikipedia Wikipedia sources seigneur **fusion** mulot **source** peuple **supprimé** habitants **références anonyme CET** archeos **articles bibliographie définition encyclopédie** huit matthieu pdd **modifs texte** panoramix

Des termes relatifs aux discussions et à la thématique

Forum Santé, F-mesure 99%

enceintes désolée **quote** poussette lila34 essais nausicaa pma bb douleur 45 zhom ui fiv grossesse ovulation **filles** écho chéri **cher** fofo dpo fc gygy récap **bisous coucou** 74 rigolo **emoticone**

N-grams de mots, WikiDiscussion vs. Forum Santé

WikiDiscussion, F-mesure 97%

l'-article cet-article bonjour-, cest-) (-cest bonjour-. un-article (-cet cet-) utilisateur-# merci-. liens-externes cordialement-. cette-page etc-. la-page des-sources cordialement-, lien-externe image-# je-propose cf-. il-faudrait

Forum Santé, F-mesure 98%

emoticone- ? emoticone-ou emoticone-de ma-courbe la-journée ton-homme emoticone-a emoticone-(emoticone- ! ! ça-va l'-ovu emoticone-pourque mon-chéri l'-ovulation les-essais emoticone-j' de-grossesse emoticone- ! ! ! ce-matin emoticone-on rigolo-# emoticone-c' emoticone- ! trop-rigolo emoticone-à comment-allez la-grossesse ma-belle coucou-les emoticone-mais mon-homme emoticone-je emoticone-pour emoticone-les emoticone-et emoticone-, emoticone-) les-filles

N-grams de mots, WikiDiscussion vs. Forum Santé

WikiDiscussion, F-mesure 97%

l'-article cet-article bonjour-, cest-) (-cest bonjour-. **un-article** (-cet cet-) **utilisateur-#** merci-. **liens-externes** cordialement-. **cette-page etc-. la-page des-sources** cordialement-, **lien-externe image-#** je-propose cf-. il-faudrait

Forum Santé, F-mesure 98%

emoticone- ? emoticone-ou emoticone-de ma-courbe la-journée ton-homme emoticone-a emoticone-(emoticone- ! ! ça-va l'-ovu emoticone-pourque mon-chéri l'-ovulation les-essais emoticone-j' de-grossesse emoticone- ! ! ! ce-matin emoticone-on rigolo-# emoticone-c' emoticone- ! trop-rigolo emoticone-à comment-allez la-grossesse ma-belle coucou-les emoticone-mais mon-homme emoticone-je emoticone-pour emoticone-les emoticone-et emoticone-, emoticone-) les-filles

N-grams de mots, WikiDiscussion vs. Forum Santé

WikiDiscussion, F-mesure 97%

l'-article cet-article **bonjour-,** cest-) (-cest **bonjour-.** un-article (-cet cet-) utilisateur-# **merci-.** liens-externes **cordialement-.** cette-page etc-. la-page des-sources **cordialement-,** lien-externe image-# je-propose cf-. il-faudrait

Forum Santé, F-mesure 98%

emoticone- ? emoticone-ou emoticone-de ma-courbe la-journée ton-homme emoticone-a emoticone-(emoticone- ! ! ça-va l'-ovu emoticone-pourque mon-chéri l'-ovulation les-essais emoticone-j' de-grossesse emoticone- ! ! ! ce-matin emoticone-on rigolo-# emoticone-c' emoticone- ! trop-rigolo emoticone-à comment-allez la-grossesse ma-belle coucou-les emoticone-mais mon-homme emoticone-je emoticone-pour emoticone-les emoticone-et emoticone-, emoticone-) les-filles

N-grams de mots, WikiDiscussion vs. Forum Santé

WikiDiscussion, F-mesure 97%

l'-article cet-article bonjour-, cest-) (-cest bonjour-. un-article (-cet cet-) utilisateur-# merci-. liens-externes cordialement-. cette-page etc-. la-page des-sources cordialement-, lien-externe image-# **je-propose cf-. il-faudrait**

Forum Santé, F-mesure 98%

emoticone- ? emoticone-ou emoticone-de ma-courbe la-journée ton-homme emoticone-a emoticone-(emoticone- ! ! ça-va l'-ovu emoticone-pourque mon-chéri l'-ovulation les-essais emoticone-j' de-grossesse emoticone- ! ! ! ce-matin emoticone-on rigolo-# emoticone-c' emoticone- ! trop-rigolo emoticone-à comment-allez la-grossesse ma-belle coucou-les emoticone-mais mon-homme emoticone-je emoticone-pour emoticone-les emoticone-et emoticone-, emoticone-) les-filles

N-grams de mots, WikiDiscussion vs. Forum Santé

WikiDiscussion, F-mesure 97%

l'-article cet-article bonjour-, cest-) (-cest bonjour-. un-article (-cet cet-) utilisateur-# merci-. liens-externes cordialement-. cette-page etc-. la-page des-sources cordialement-, lien-externe image-# je-propose cf-. il-faudrait

Forum Santé, F-mesure 98%

emoticone- ? emoticone-ou emoticone-de ma-courbe la-journée ton-homme **emoticone-a emoticone-(emoticone- ! !** ça-va l'-ovu **emoticone-pourque** mon-chéri l'-ovulation les-essais **emoticone-j'** de-grossesse **emoticone- ! ! ! ce-matin emoticone-on** rigolo-# **emoticone-c' emoticone- !** trop-rigolo **emoticone-à** comment-allez la-grossesse ma-belle coucou-les **emoticone-mais** mon-homme **emoticone-je emoticone-pour emoticone-les emoticone-et emoticone-, emoticone-)** les-filles

N-grams de mots, WikiDiscussion vs. Forum Santé

WikiDiscussion, F-mesure 97%

l'-article cet-article bonjour-, cest-) (-cest bonjour-. un-article (-cet cet-) utilisateur-# merci-. liens-externes cordialement-. cette-page etc-. la-page des-sources cordialement-, lien-externe image-# je-propose cf-. il-faudrait

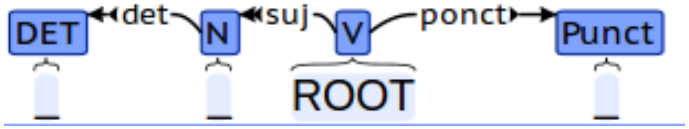
Forum Santé, F-mesure 98%

emoticone- ? emoticone-ou emoticone-de **ma-courbe la-journée ton-homme** emoticone-a emoticone-(emoticone- ! ! ça-va **l'-ovu** emoticone-pourque **mon-chéri l'-ovulation les-essais** emoticone-j' **de-grossesse** emoticone- ! ! ! ce-matin emoticone-on **rigolo-#** emoticone-c' emoticone- ! trop-rigolo emoticone-à comment-allez la-grossesse ma-belle coucou-les emoticone-mais mon-homme emoticone-je emoticone-pour emoticone-les emoticone-et emoticone-, emoticone-) les-filles

N-grams syntaxiques : tri-arcs avec DEP + POS + MORPHO

WikiDiscussion F-mesure 91%

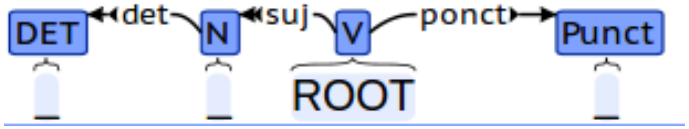
Forte présence du tri-arc (avec des traits morphologiques différents)



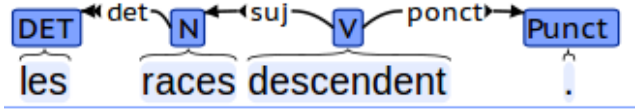
N-grams syntaxiques : tri-arcs avec DEP + POS + MORPHO

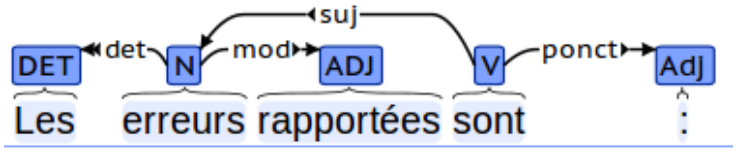
WikiDiscussion F-mesure 91%

Forte présence du tri-arc (avec des traits morphologiques différents)



Exemples de contextes qui impliquent cet N-gram :

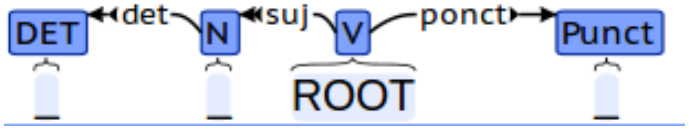




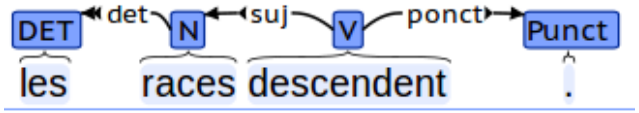
N-grams syntaxiques : tri-arcs avec DEP + POS + MORPHO

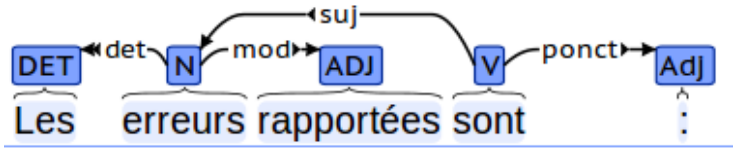
WikiDiscussion F-mesure 91%

Forte présence du tri-arc (avec des traits morphologiques différents)

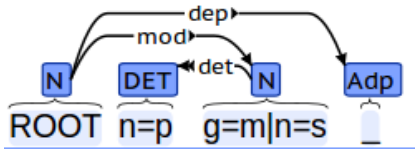


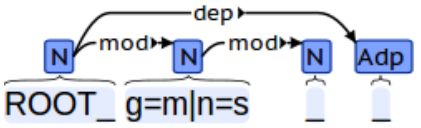
Exemples de contextes qui impliquent cet N-gram :





Signatures (code wiki) :

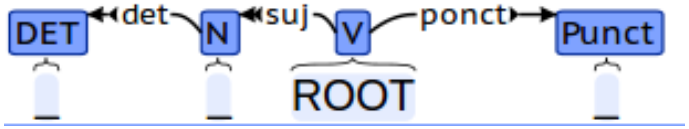




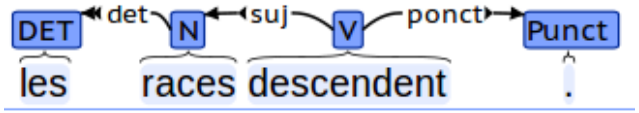
N-grams syntaxiques : tri-arcs avec DEP + POS + MORPHO

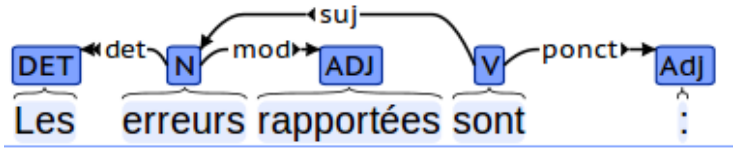
WikiDiscussion F-mesure 91%

Forte présence du tri-arc (avec des traits morphologiques différents)

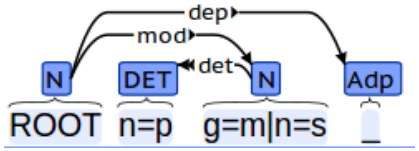


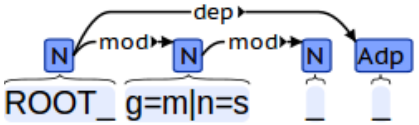
Exemples de contextes qui impliquent cet N-gram :

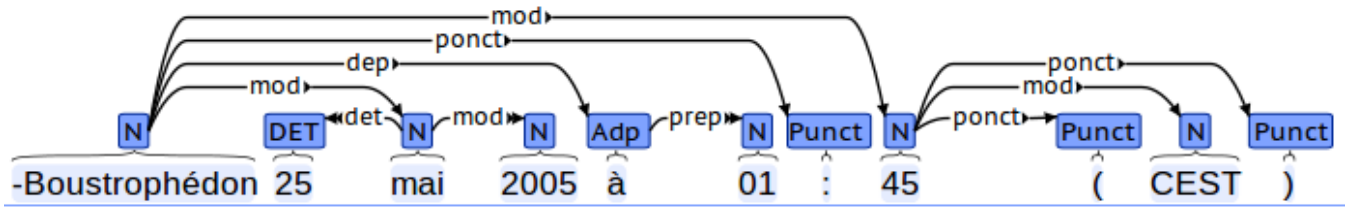




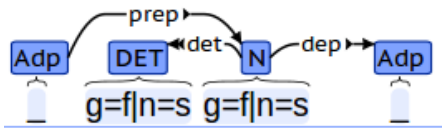
Signatures (code wiki) :

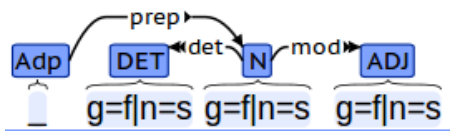




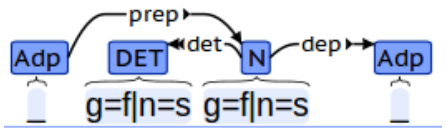


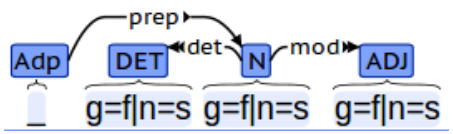
Syntagmes prépositionnelles :

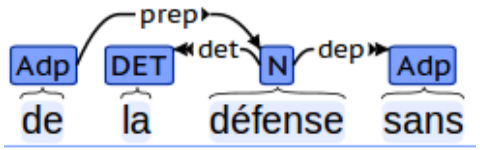


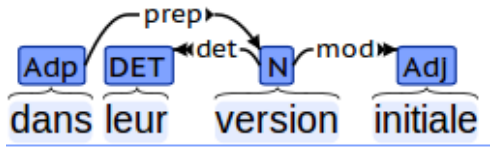


Syntagmes prépositionnelles :

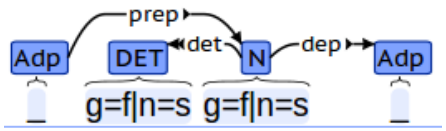


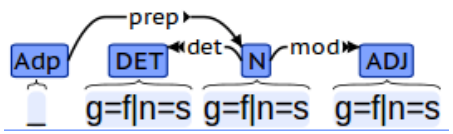


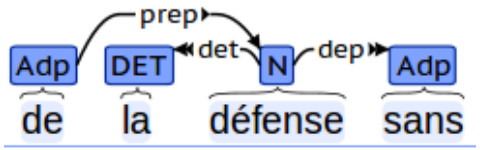


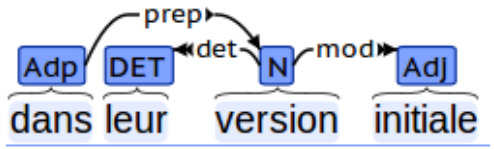


Syntagmes prépositionnelles :

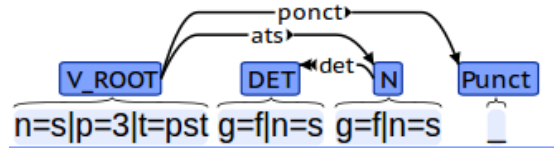




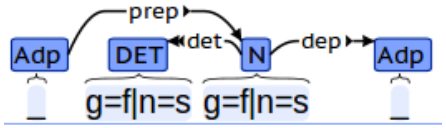


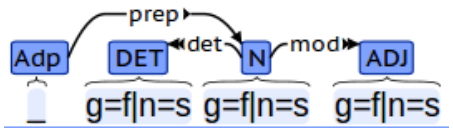


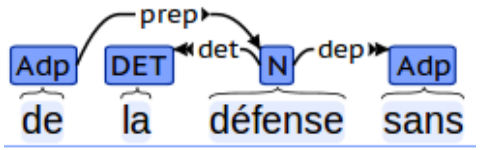
Construction *est*+NOM :

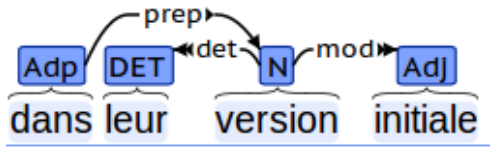


Syntagmes prépositionnelles :

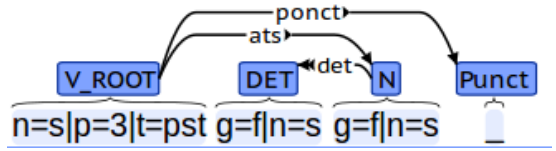


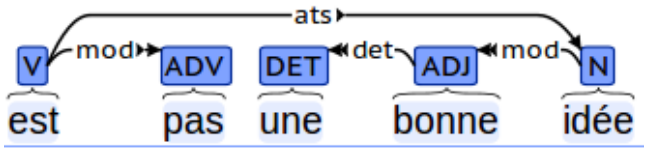


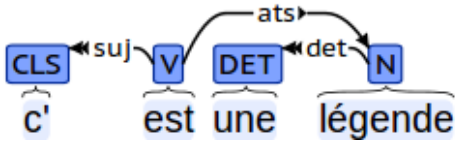




Construction *est*+NOM :

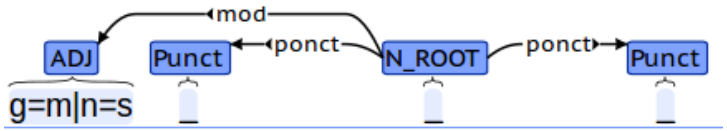






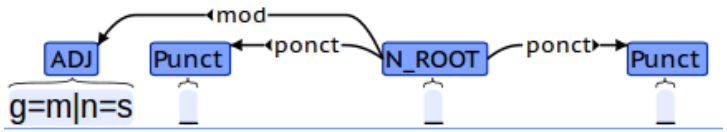
Forum Santé F-mesure 94%

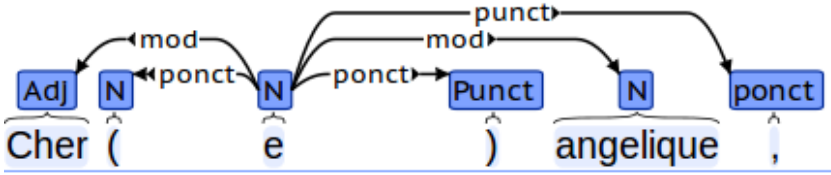
Salutations :



Forum Santé F-mesure 94%

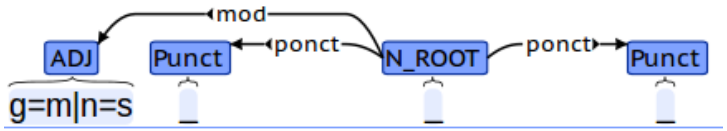
Salutations :

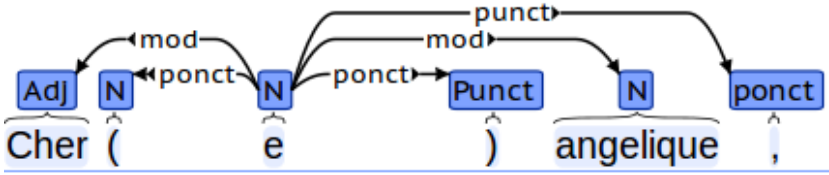




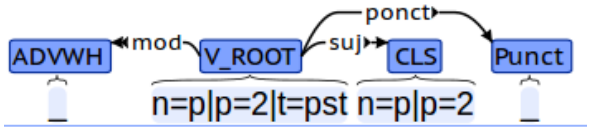
Forum Santé F-mesure 94%

Salutations :



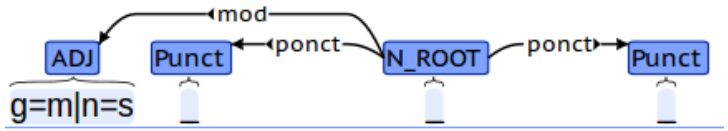


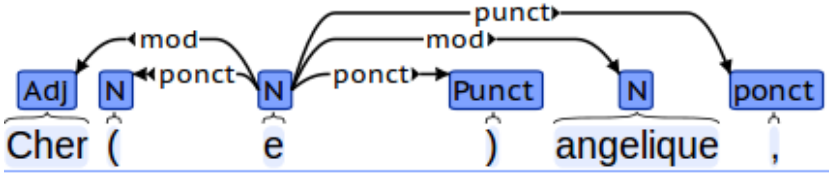
Questions :



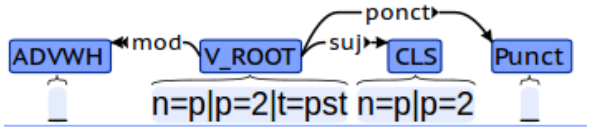
Forum Santé F-mesure 94%

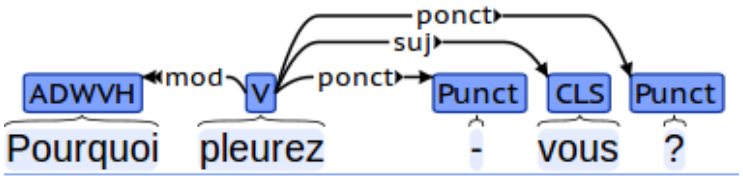
Salutations :



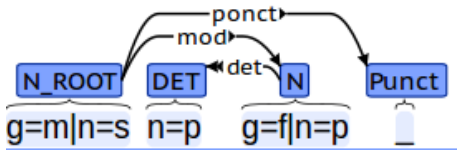


Questions :

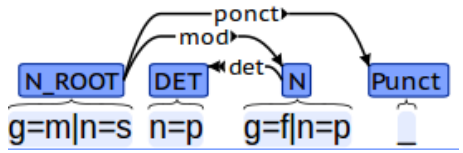


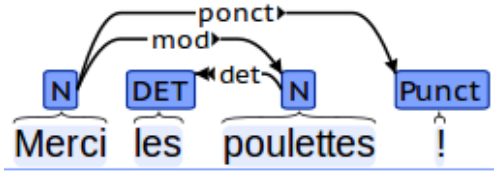


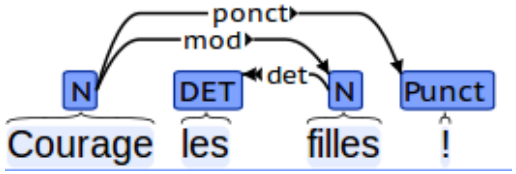
Souhaits :



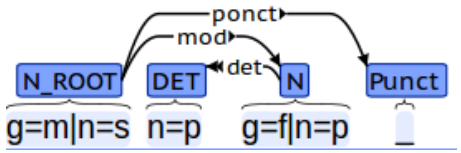
Souhaits :

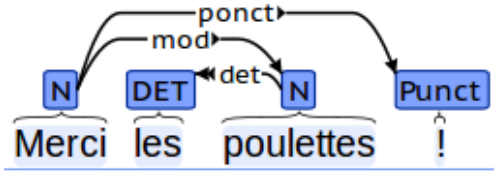

ponct
mod
det
N_ROOT
DET
N
Punct
g=m|n=s
n=p
g=f|n=p
_

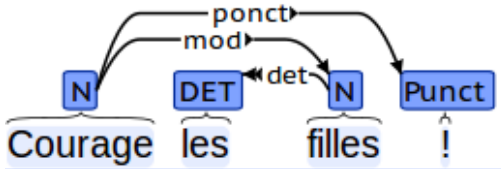

ponct
mod
det
N
DET
N
Punct
Merci
les
poulettes
!


ponct
mod
det
N
DET
N
Punct
Courage
les
filles
!

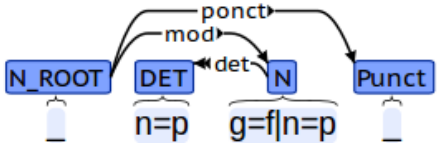
Souhaits :



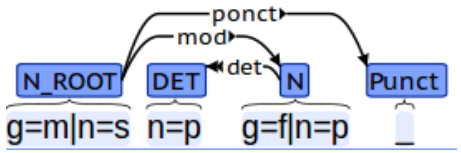


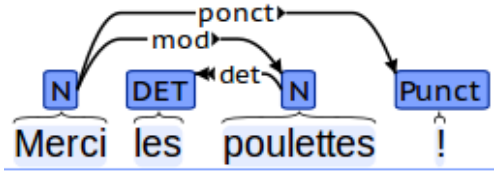


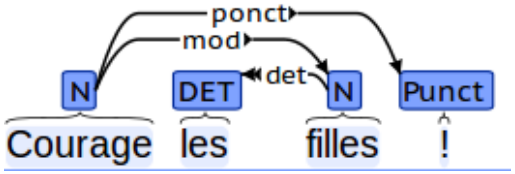
Émoticones :



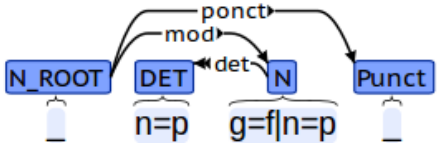
Souhaits :

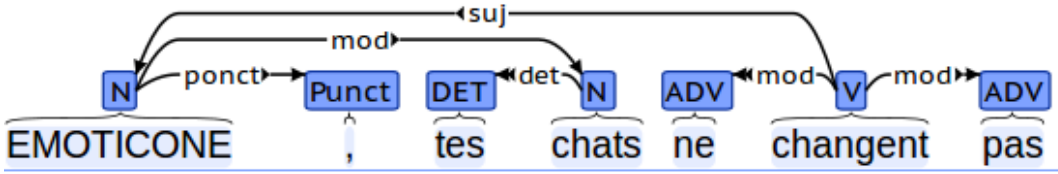






Émoticones :





Conclusion sur WikiDiscussion vs. Forum Santé

- Les deux corpus montrent des caractéristiques communes : noms d'utilisateurs, phrases nominales (typiques des discussions ?)
- ... mais on voit également des différences très claires !
- Les N-grams des WikiDiscussion sont associés à des phrases complètes (même complexes)
- .... tandis que les N-grams du Forum de Santé sont associés à des souhaits, des salutations familières et des emoticones

bag-of-lemma WikiDiscussion vs. articles Wikipedia

WikiDiscussion, F-mesure 92%

bonjour cordialement merci mar paragraphe scientologie info wikipedia article sep jan oeuvre d'accord wallon bandeau bonsoir ça

articles Wikipedia, F-mesure 92%

mikhaïl précipitation posthume comète subdivision équiper skinhead polygone réalisateur hugues console scrutin accusatif verne horus millimètre poète

bag-of-lemma WikiDiscussion vs. articles Wikipedia

WikiDiscussion, F-mesure 92%

bonjour cordialement merci mar paragraphe scientologie info wikipedia article sep jan oeuvre d'accord wallon bandeau **bonsoir** ça

articles Wikipedia, F-mesure 92%

mikhaïl précipitation posthume comète subdivision équiper skinhead polygone réalisateur hugues console scrutin accusatif verne horus millimètre poète

bag-of-lemma WikiDiscussion vs. articles Wikipedia

WikiDiscussion, F-mesure 92%

bonjour cordialement merci mar **paragraphe** scientologie **info wikipedia article** sep jan oeuvre d'accord wallon bandeau bonsoir ça

articles Wikipedia, F-mesure 92%

mikhaïl précipitation posthume comète subdivision équiper skinhead polygone réalisateur hugues console scrutin accusatif verne horus millimètre poète

bag-of-lemma WikiDiscussion vs. articles Wikipedia

WikiDiscussion, F-mesure 92%

bonjour cordialement merci mar paragraphe **scientologie** info wikipedia article sep jan oeuvre d'accord **wallon bandeau** bonsoir ça

articles Wikipedia, F-mesure 92%

mikhaïl précipitation posthume comète subdivision équiper skinhead polygone réalisateur hugues console scrutin accusatif verne horus millimètre poète

bag-of-lemma mais Big(ger) Data – 12 million de mots

WikiDiscussion, F-Score 92% → 93%

bonjour cordialement merci mar article oeuvre discuter paragraphe info wikipedia bonsoir bandeau je ça sep jan avis

⇒ **Peu de différences en terme de F-mesure et de lemmes typiques**

articles Wikipedia, F-mesure 92% → 93%

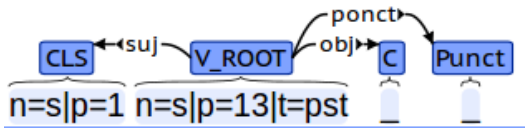
island console shogun réalisateur polygone zelda poète astrologie échiquier saints manga jonathan matador député maire retenue footballeur dicton

⇒ **Peu de différences en terme de F-mesure et mais des différences en terme de lemmes typiques**

⇒ **Limite de l'approche lexicale**

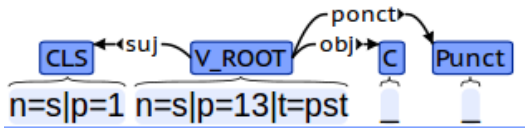
Wikipedia discussions, F-mesure 86%

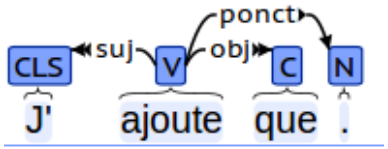
Je+verbe+que :

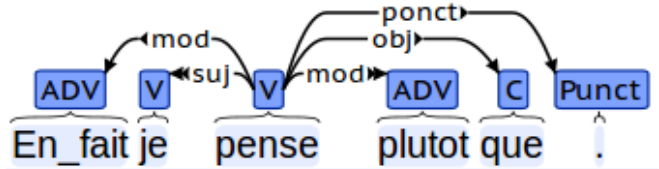


Wikipedia discussions, F-mesure 86%

Je+verbe+que :

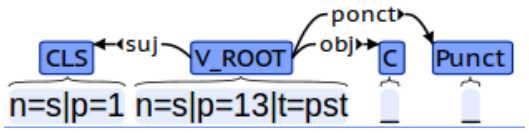


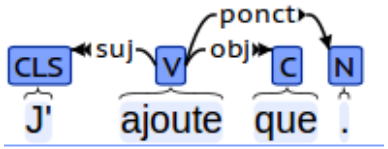


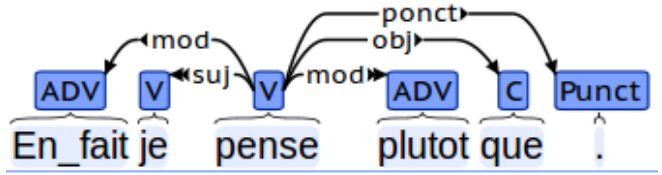


Wikipedia discussions, F-mesure 86%

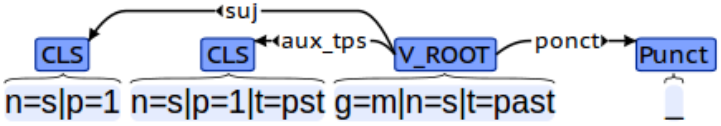
Je+verbe+que :





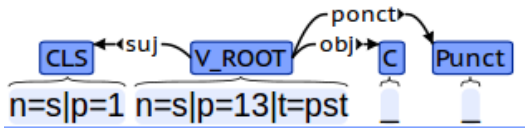


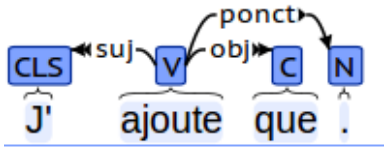
Je+aux+verbe :

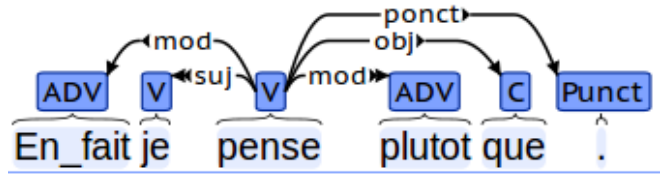


Wikipedia discussions, F-mesure 86%

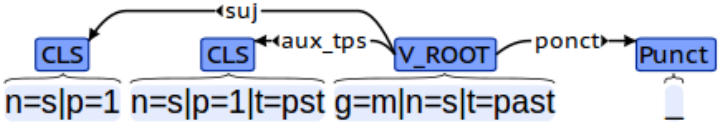
Je+verbe+que :

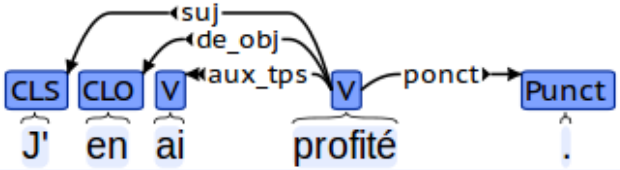


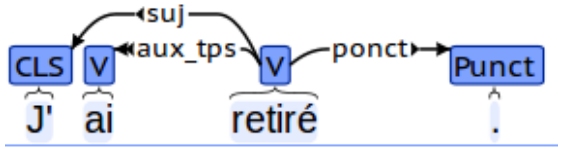




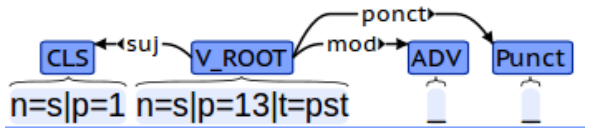
Je+aux+verbe :



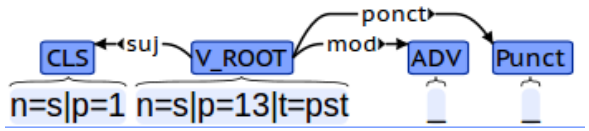


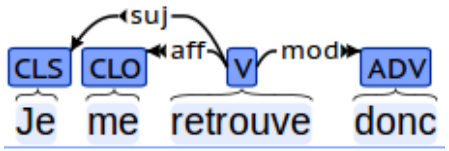


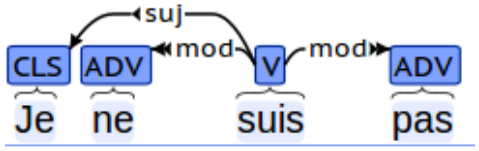
Je+verbe+ADV :

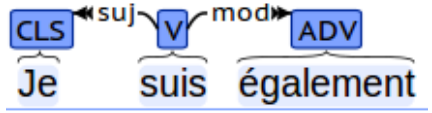


Je+verbe+ADV :

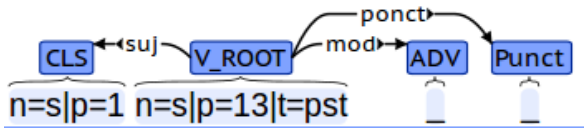

CLS
suj
V_ROOT
ponct
mod
ADV
Punct
n=s|p=1
n=s|p=13|t=pst
_
_

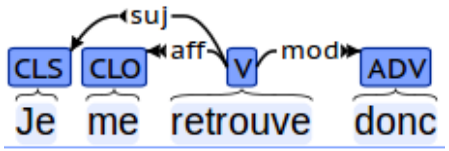

suj
CLS
CLO
aff
V
mod
ADV
Je
me
retrouve
donc

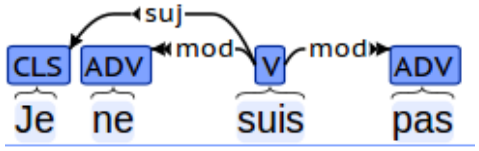

suj
CLS
ADV
mod
V
mod
ADV
Je
ne
suis
pas

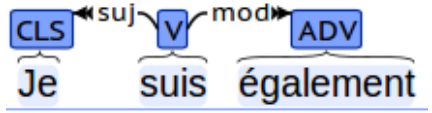

CLS
suj
V
mod
ADV
Je
suis
également

Je+verbe+ADV :

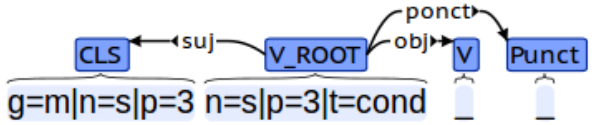




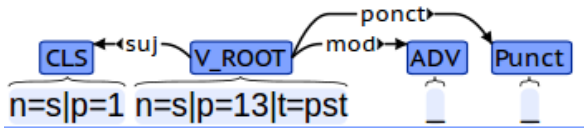


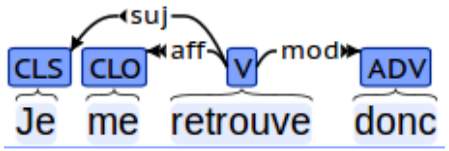


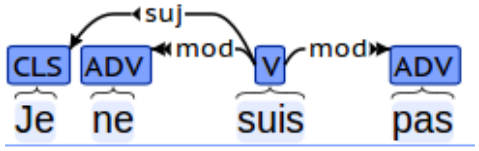
CLS + cond. → Opinions ! :

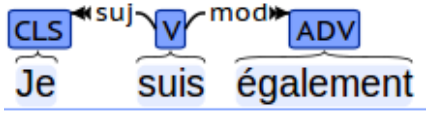


Je+verbe+ADV :

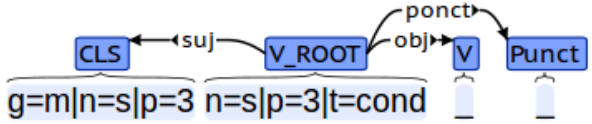


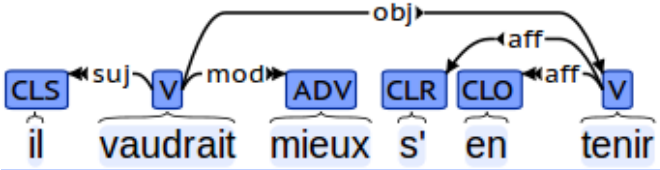


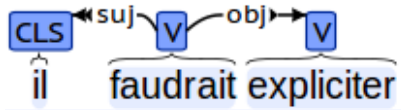




CLS + cond. → Opinions ! :

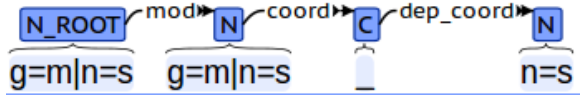






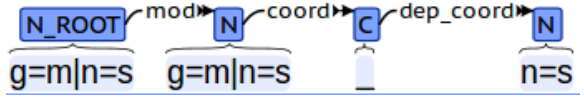
articles Wikipedia, F-mesure 85%

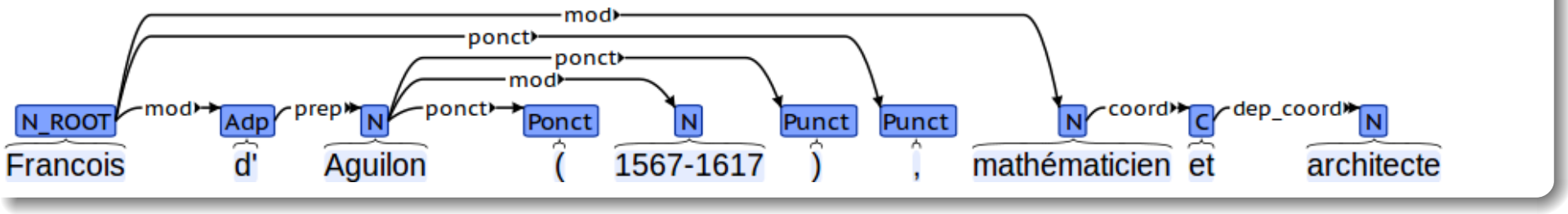
Coordinations dans une phrase nominale :



articles Wikipedia, F-mesure 85%

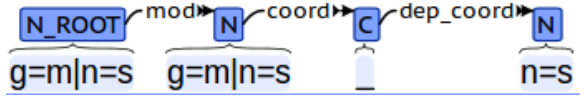
Coordinations dans une phrase nominale :

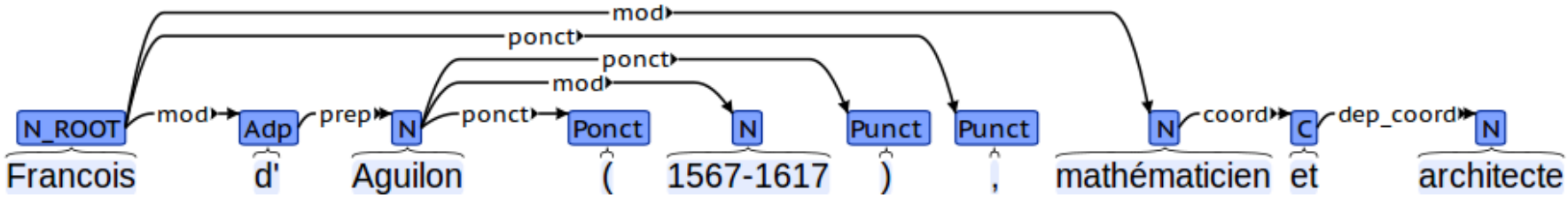




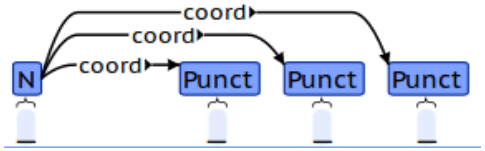
articles Wikipedia, F-mesure 85%

Coordinations dans une phrase nominale :



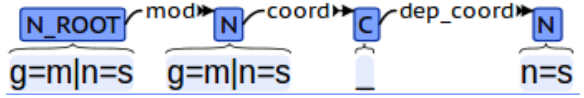


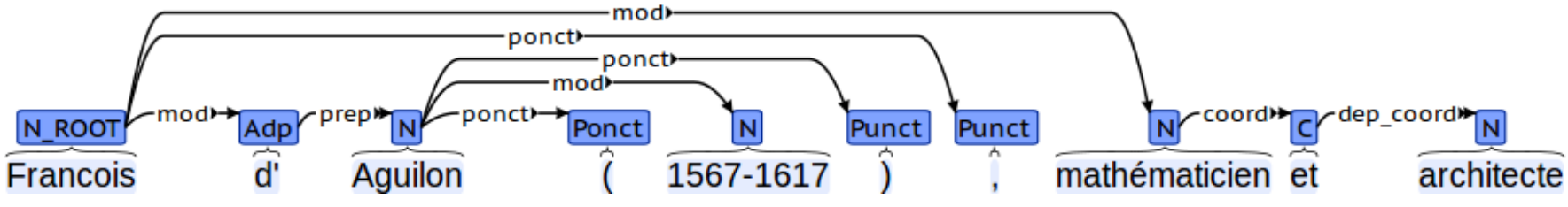
Autres coordinations :



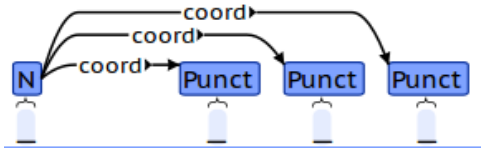
articles Wikipedia, F-mesure 85%

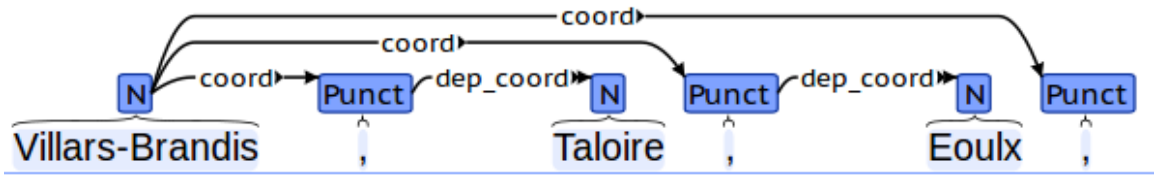
Coordinations dans une phrase nominale :



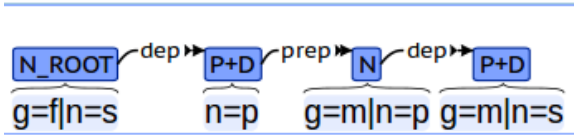


Autres coordinations :

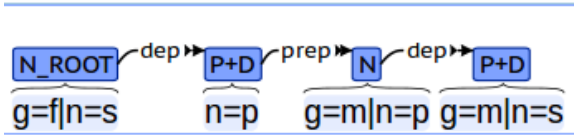


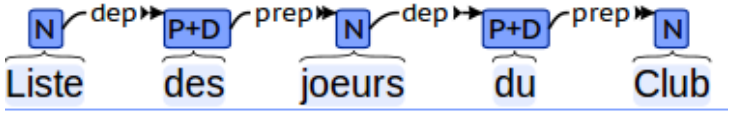


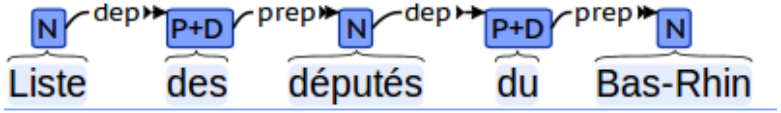
Phrases nominales annoncant le début d'une liste :



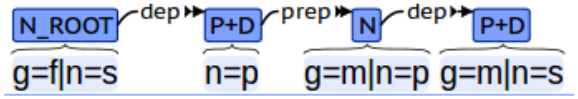
Phrases nominales annoncant le début d'une liste :

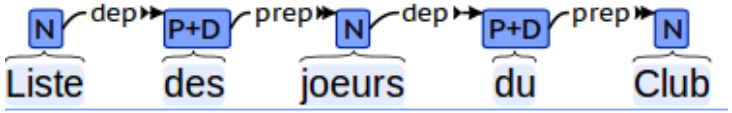


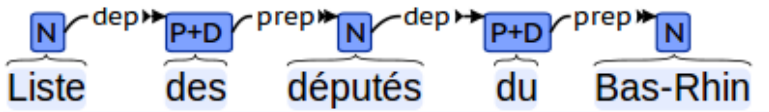




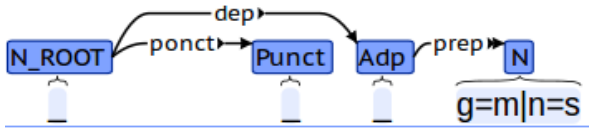
Phrases nominales annoncant le début d'une liste :



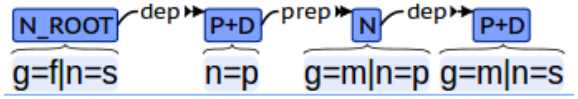


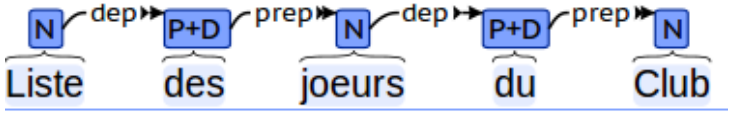


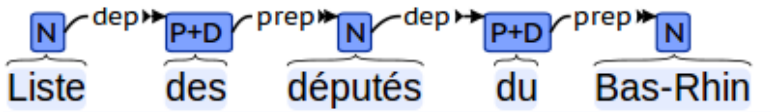
Phrases nominales explicatives :



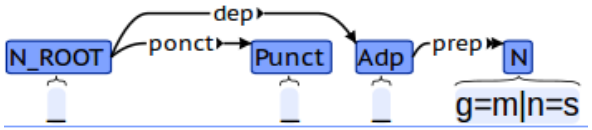
Phrases nominales annoncant le début d'une liste :

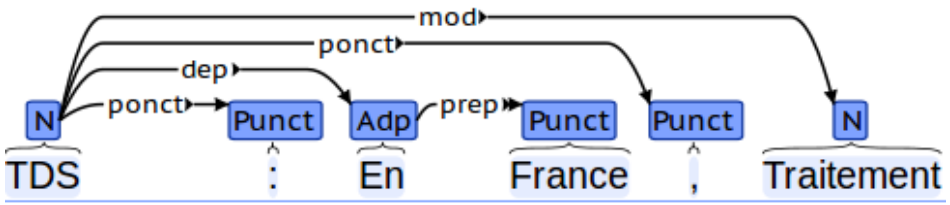






Phrases nominales explicatives :





Résultats

WikiDiscussion ↔ Articles Wikipedia

- WikiDiscussion : combinaisons différentes de *je* + *verbe* et *il* + *cond.*
  - → Modalité / subjectivité / argumentation (à creuser)
- Articles Wikipedia : coordinations, phrases nominales
  - → Annonces des débuts des listes, des listes
  - → Constructions visant à expliquer + informer

WikiDiscussion ↔ Forum Santé

- WikiDiscussion : combinaisons différentes de *det+suj+verbe*, phrases prépositionnelles et *est* + *verbe*
  - → Phrases complètes
- Forum Santé : souhaits, salutations, emoticones

Caractéristiques des WikiDiscussions basées sur les N-grams syntaxiques

- Pas de traits de langage familier !
- Pas de traits de signes d'interaction
- N-grams syntaxiques reflètent plutôt une syntaxe complète
- Traces de subjectivité / argumentation

Et l'utilité des n-grams syntaxiques ?

Petite comparaison aux trigrammes lexicaux

WikiDiscussions, F-mesure 88%

(-cest-) (-cet-) "'-pierre-"' de-l'-article dans-l'-article **il-me-semble ,-j'-ai** l'-article-. **je-n'-ai je-pense-que que-j'-ai ,-je-ne** dans-cet-article je-viens-de de-la-page c'-est-pas que-c'-est **je-ne-sais** il-y-a cet-article-. **je-ne-vois je-pense-qu'** lien-externe-mort **je-suis-d'accord ,-non- ?** je-l'-ai ,-c'-est

Articles Wikipedia F-mesure 89%

avions-actuels-# (-né-en (-ambassade-) ,-gallimard-, danseur-et-chorégraphe (-bretagne-) selon-la-liste artiste-professionnel-# ,-né-à festival-de-cannes)-,-peintre (-danseur-et écrivain-amateur-# liste-des-sénateurs communauté-de-communes)—royaume-uni liste-des-députés (-éteint-))-,-écrivain)-,-français prix-campbell-# ,-coll-. liste-des-préfets artiste-amateur-# #-ee-siècle

N-grams syntaxiques $\leftrightarrow$ lexicaux

WikiDiscussion

- Présence des traces de subjectivité avec les deux méthodes
- Présence des phrases complètes uniquement avec les N-grams syntaxiques

Articles de Wikipedia

- Présence des coordinations, listes, explications uniquement avec les N-grams syntaxiques

Bilan de l'approche par N-grams syntaxique

Bilan des F-mesures

	Wikidiscussions	**Forum Santé**
Bag-of-words	100%	99%
N-grams syntaxiques	91%	94%
	Wikidiscussions	**Articles Wikipédia**
Bag-of-lemmata	92%	92%
N-grams syntaxiques	86%	85%
Trigrammes lexicaux	88%	89%

Conclusions

- → Les N-grams syntaxiques moyen de classification relativement efficace
- → Certaines caractéristiques reflétées par des N-grams syntaxiques aussi présentes dans les N-grams lexicaux
- → Certaines caractéristiques structurales complètement absentes des N-grams lexicaux

Plan

1 Motivations : Web As Corpus

2 Wikipedia As Corpus

3 Premières analyses

4 Conclusions et perspectives

Conclusions générales

Sur la caractérisation des WikiDiscussions

- Des discussions "bien écrites" (phrases complètes)
- Un niveau de langue standard (non familier)
- A priori peu d'*affects* mais de la modalisation et de l'implication du locuteur (je+V)
- L'objectif d'un consensus

Sur la complémentarité des approches

- Les résultats semblent se compléter

To Dos

- Optimisation des paramètres pour pouvoir utiliser des corpus plus grands
  - Sélection des traits (en relation avec les analyses hypothesis-driven)
  - Sélection des discussions "longues" pour l'apprentissage
  - Taille des segments à classifier : message entier / paragraphe d'article (permet l'accès aux débuts et fins)
- Application de la méthode pour un corpus avec une autre thématique (out of domain) pour examiner son adaptabilité
  - Sudiviser le corpus et tirer partie des méta-données (portail physique vs. portail people)
  - Contraster avec un forum de discussion différent (forum Ubuntu ?)
- Classification non supervisée
- Prolonger les études sur des marqueurs spécifiques : noms sous-spécifiés et vocabulaire familier

Références

- Agarwal, A., Chappelle, O., Dudik, M. and Langford, J. (2011). A Reliable Effective Terascale Linear Learning System. JMLR, 15, 1111-1133.
- Augustyn, M., Ben Hamou, S., Bloquet, G., Goossens, V., Loiseau, M., & Rynck, F. (2008) Constitution de ressources pédagogiques numériques : le lexique des affects. Dans M. Loiseau, M. Abouzaïd, L. Buson, C. Cavalla, A. Djaroun, C. Dugua, et al. (éD.), Autour Des Langues Et Du Langage : Perspective Pluridisciplinaire (p. 407–414). Grenoble : Presses Universitaires de Grenoble.
- Goldberg, Y., and Orwant, J. (2013). A Dataset of Syntactic-N grams over Time from a Very Large Corpus of English Books. Second Joint Conference on Lexical and Computational Semantics (*SEM), 1. Association for Computational Linguistics.
- Kanerva, J., Luotolahti, J., Laippala, V., and Ginter, F. (2014). Syntactic N-gram Collection from a Large-Scale Corpus of Internet Finnish. Proceedings of the Sixth International Conference Baltic HLT.
- Kilgarriff, A., Reddy, S., Pomikálek, J. and PVS A. (2010). A Corpus Factory for Many Languages. In LREC workshop on Web Services and Processing Pipelines, Malta, May 2010.
- Laippala, Veronika ; Kanerva, Jenna ; Ginter, Filip. 2015. Syntactic Ngrams as Keystructures Reflecting Typical Syntactic Patterns of Corpora in Finnish. Procedia – Social and Behavioral Sciences. Current Work in Corpus Linguistics. 198, 233-241.
- Laippala, Veronika ; Kanerva, Jenna ; Pyysalo, Sampo ; Missilä, Anna ; Salakoski, Tapio ; Ginter, Filip. 2015. Syntactic N-grams in the Classification of the Finnish Internet Parsebank : Detecting Translations and Informality. Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015), May 11–13, 2015 in Vilnius, Lithuania.
- Scott, M., and Tribble, C. (2006). Textual Patterns : Key Words and Corpus Analysis in Language Education . Philadelphia, PA, USA : John Benjamins Publishing Company.