Genres textuels et caractéristiques stylistiques pour la classification

gael.lejeune@sorbonne-universite.fr

24 avril 2025

Sorbonne Université

Les propriétés du genre textuel

Summary of ChatGPT-Related research and perspective towards the future of large language models

Yiheng Liu [a,1], Tianle Han [a,1], Siyuan Ma [a], Jiayue Zhang [a], Yuanyuan Yang [a], Jiaming Tian [a], Hao He [a], Antong Li [b], Mengshen He [a], Zhengliang Liu [c], Zihao Wu [c], Lin Zhao [c], Dajiang Zhu [d], Xiang Li [e], Ning Qiang [a], Dingang Shen [f,g,h], Tianming Liu [c], Bao Ge [a,*]

[a] School of Physics and Information Technology, Shaanxi Normal University, Xi'an, 710119, Shaanxi, China
[b] School of Life and Technology Biomedical-Engineering, Xi'an Jiaotong University, Xi'an, 710119, Shaanxi, China
[c] School of Computing, The University of Georgia, Athens, 30602, USA
[d] Department of Computer Science and Engineering, The University of Texas at Arlington, Arlington, 76019, USA
[e] Department of Radiology, Massachusetts General Hospital and Harvard Medical School, Boston, 02115, USA
[f] School of Biomedical Engineering, ShanghaiTech University, Shanghai, 201210, China
[g] Shanghai United Imaging Intelligence Co., Ltd., Shanghai, 200230, China
[h] Shanghai Clinical Research and Trial Center, Shanghai, 201210, China

ABSTRACT

This paper presents a comprehensive survey of ChatGPT-related (GPT-3.5 and GPT-4) research, state-of-the-art large language models (LLM) from the GPT series, and their prospective applications across diverse domains. Indeed, key innovations such as large-scale pre-training that captures knowledge across the entire world wide web, instruction fine-tuning and Reinforcement Learning from Human Feedback (RLHF) have played significant roles in enhancing LLMs' adaptability and performance. We performed an in-depth analysis of 194 relevant papers on arXiv, encompassing trend analysis, word cloud representation, and distribution analysis across various application domains. The findings reveal a significant and increasing interest in ChatGPT-related research, predominantly centered on direct natural language processing applications, while also demonstrating considerable potential in areas ranging from education and history to mathematics, medicine, and physics. This study endeavors to furnish insights into ChatGPT's capabilities, potential implications, ethical concerns, and offer direction for future advancements in this field.

1. Introduction

Recent advances in natural language processing (NLP) have led to the development of powerful language models such as the Generative Pre-trained Transformer (GPT) series,[1,2,3–5] including large language models (LLM) such as ChatGPT (GPT-3.5 and GPT-4).[6] These models are pre-trained on vast amounts of text data and have demonstrated exceptional performance in a wide range of NLP tasks, including language translation, text summarization, and question-answering. In particular, the ChatGPT model has demonstrated its potential in various fields, including education, healthcare, reasoning, text generation, human-machine interaction, and scientific research.

A key milestone of LLM development is InstructGPT,[2] a framework that allows for instruction fine-tuning of a pre-trained language model based on reinforcement learning from human feedback (RLHF).[7,2] This framework enables an LLM to adapt to a wide range of NLP tasks, making it highly versatile and flexible by leveraging human feedback. RLHF enables the model to align with human preferences and human values, which significantly improves from large language models that are solely trained text corpora through unsupervised pre-training. ChatGPT is a successor to InstructGPT. Since its release in December 2022, ChatGPT has been equipped with these advanced developments, leading to impressive performances in various downstream NLP tasks such as reasoning and generalized text generation. These unprecedented NLP capabilities spur applications in diverse domains such as education, healthcare, human-machine interaction, medicine and scientific research. ChatGPT has received widespread attention and interest, leading to an increasing number of applications and research that harness its exceeding potential.

The open release of the multi-modal GPT-4 model further expands the horizon of large language models and empowers exciting developments that involve diverse data beyond text.

The purpose of this paper is to provide a comprehensive survey of the existing research on ChatGPT and its potential applications in various

* Corresponding author.
E-mail address: bob_ge@snnu.edu.cn (B. Ge).
[1] These authors contributed equally to this work.

SCIENCES • INTELLIGENCE ARTIFICIELLE

Les défauts de ChatGPT sous la loupe des scientifiques

CHRONIQUE

David Larousserie

Une équipe de chercheurs sino-américains a évalué ce nouveau robot conversationnel en étudiant 194 prépublications (« preprints ») sur près de 1 400, déposées depuis décembre 2022. Résultat : la liste des imperfections intrinsèques de la machine est longue.

Publié le 18 octobre 2023 à 18h00 | Lecture 2 min.

Ajouter à vos sélections

Article réservé aux abonnés

Vie des labos. Il n'y a pas que les médias, les étudiants ou les artistes qui se passionnent pour ChatGPT, l'agent conversationnel de l'entreprise américaine OpenAI, qui a réveillé l'intérêt pour l'intelligence artificielle. Les scientifiques eux-mêmes se sont penchés sur le nouveau venu, pour l'étudier sous toutes les coutures. Près de 1 400 prépublications (*preprints*) ont été déposées sur le principal site, Arxiv.org, depuis le 30 novembre 2022 et la mise en ligne de ChatGPT. La première dès le 12 décembre (sur l'analyse des premiers tweets sur le sujet)...

Un article scientifique sur Chat GPT, et l'article du Monde qui en parle

Adapter la méthode au genre textuel
Extraction de Structure (E. Giguet GREYC, et A. Barbaresi BBAW)

Robustesse du genre/style à la variation en langue

В Минздрав передали результаты исследований обновленного «Спутника»

Commons.wikimedia.org

Москва, 6 декабря - АиФ-Москва.

В регистрационное досье вакцины от COVID-19 «Спутник V» внесут изменения в связи с обновлением ее состава. С этой целью в министерство здравоохранения направлены результаты клинических исследований обновленного препарата, сообщает РИА Новости со ссылкой на директора НИЦЭМ им. Гамалеи Александра Гинцбурга.

Covid-19 : un million de doses du vaccin russe livrées à Gaza depuis les Émirats

Par Le Figaro avec AFP
Publié le 26/01/2022 à 13:40

Copier le lien

Les doses seront destinées en priorité aux écoliers de plus de 12 ans. *DADO RUVIC / REUTERS*

Un million de doses du vaccin anticoronavirus russe Spoutnik ont été livrées mercredi 26 janvier à la bande de Gaza, un don des Émirats arabes unis à l'enclave palestinienne confrontée à une hausse de contaminations par le variant Omicron.

Deux articles sur le vaccin sputnik (russe et français)

Valoriser les propriétés du matériau
Multilingue (Thèse de S. Mutuvi)

Variations et influence sur les méthodes

Gérer les sensibilités

Verrous	Solutions ?
Sensibilité à l'historicité	`modernisation`
(in-)Sensibilité au genre textuel	approche `phrastique`
Sensibilité à la langue	`Large Language Models`
Sensibilité au bruit	`correction automatique`

Variations et influence sur les méthodes

Gérer les sensibilités

Verrous	**Solutions ?**
Sensibilité à l'historicité	`modernisation`
(in-)Sensibilité au genre textuel	approche `phrastique`
Sensibilité à la langue	`Large Language Models`
Sensibilité au bruit	`correction automatique`
Toute sensibilité...	Hum, `Large Language Models` ?

Variations et influence sur les méthodes

Gérer les sensibilités

Verrous	**Solutions ?**
Sensibilité à l'historicité	`modernisation`
(in-)Sensibilité au genre textuel	approche `phrastique`
Sensibilité à la langue	`Large Language Models`
Sensibilité au bruit	`correction automatique`
Toute sensibilité...	Hum, `Large Language Models` ?

Des solutions réductionnistes :

- Réduire les documents à du "texte" (sac de phrases ?)

Variations et influence sur les méthodes

Gérer les sensibilités

Verrous	**Solutions ?**
Sensibilité à l'historicité	`modernisation`
(in-)Sensibilité au genre textuel	approche `phrastique`
Sensibilité à la langue	`Large Language Models`
Sensibilité au bruit	`correction automatique`
Toute sensibilité...	Hum, `Large Language Models` ?

Des solutions réductionnistes :

- Réduire les documents à du "texte" (sac de phrases ?)
- Faire converger ces données vers un état connu (un standard)

Faut-il toujours standardiser ?

Données Tout Gallica 1600-1800 [1]

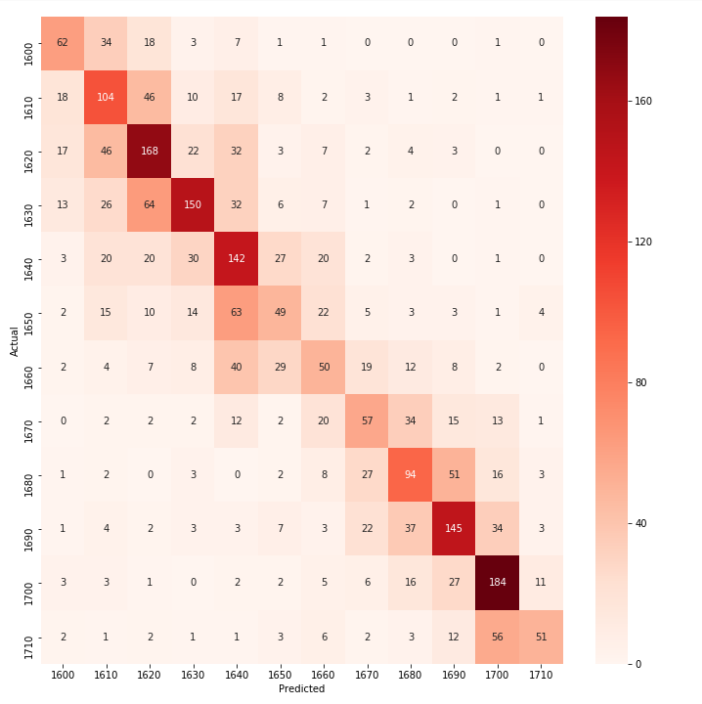
Tâche Datation automatique (en contexte bruité)

1. [Baledent et al., 2020]

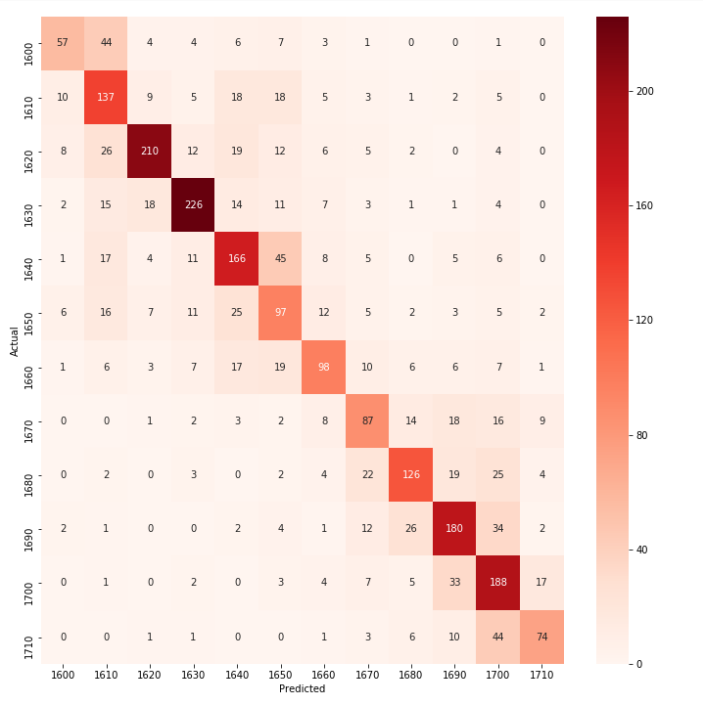
Faut-il toujours standardiser ?

Données Tout Gallica 1600-1800 [1]

Tâche Datation automatique (en contexte bruité)



Mots (F.mes 46.3, Sim. 92.8)



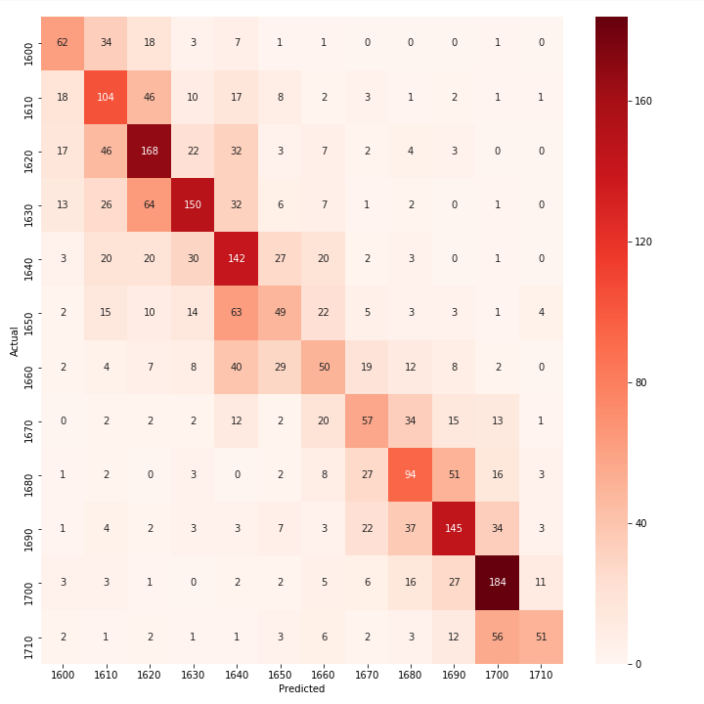
1 — 4 grammes (F-mes 71.43, Sim. 95.0)

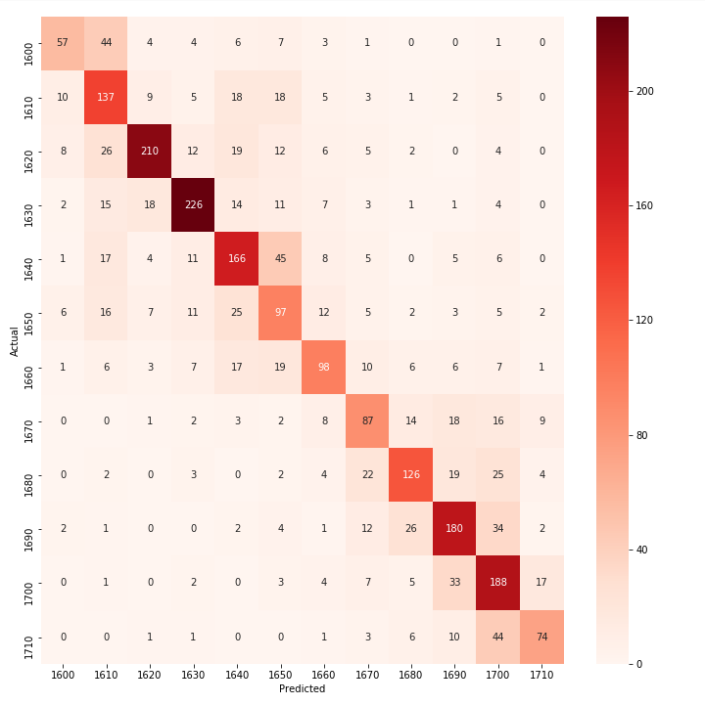
1. [Baledent et al., 2020]

Faut-il toujours standardiser ?

Données Tout Gallica 1600-1800 [1]

Tâche Datation automatique (en contexte bruité)





Mots (F.mes 46.3, Sim. 92.8)

1 – 4 grammes (F-mes 71.43, Sim. 95.0)

→ Au grain corpus, on peut s'accomoder du bruit

1. [Baledent et al., 2020]

Faut-il tout traiter au grain phrase ?

Données Corpus multilingue bruité artificiellement [2]

Tâche Classification et détection d'évènements

2. [Nguyen et al., 2020]

Faut-il tout traiter au grain phrase ?

Données Corpus multilingue bruité artificiellement [2]

Tâche Classification et détection d'évènements

(i)

(ii)

(iii)

(iv)

(v)

(vi)

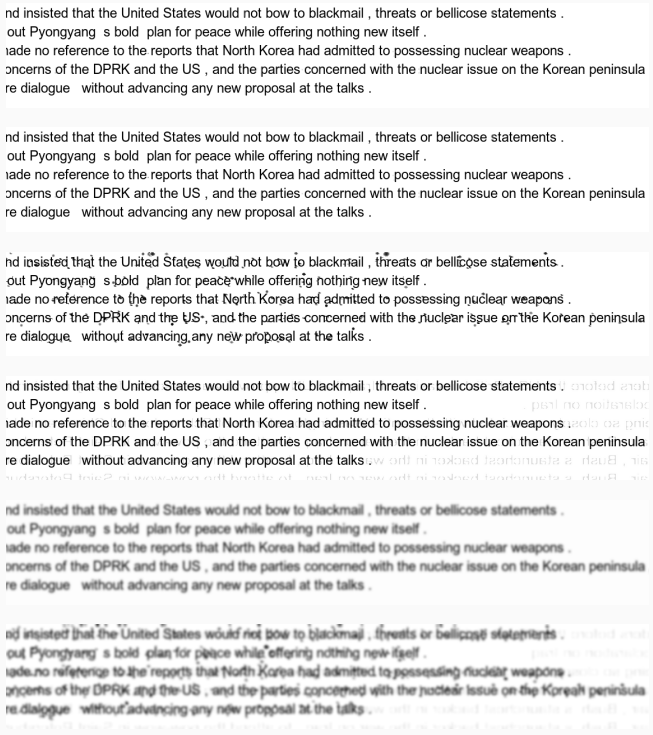


Figure 4 – (i) clean (ii) *Phantom*, (iii) *CharDeg*, (iv) *Bleed*, (v) *Blur*, (vi) all.

→ **au grain document, on peut s'accomoder du bruit**

2. [Nguyen et al., 2020]

Faut-il purger les XML ?

Données Corpus XML annoté terminologiquement

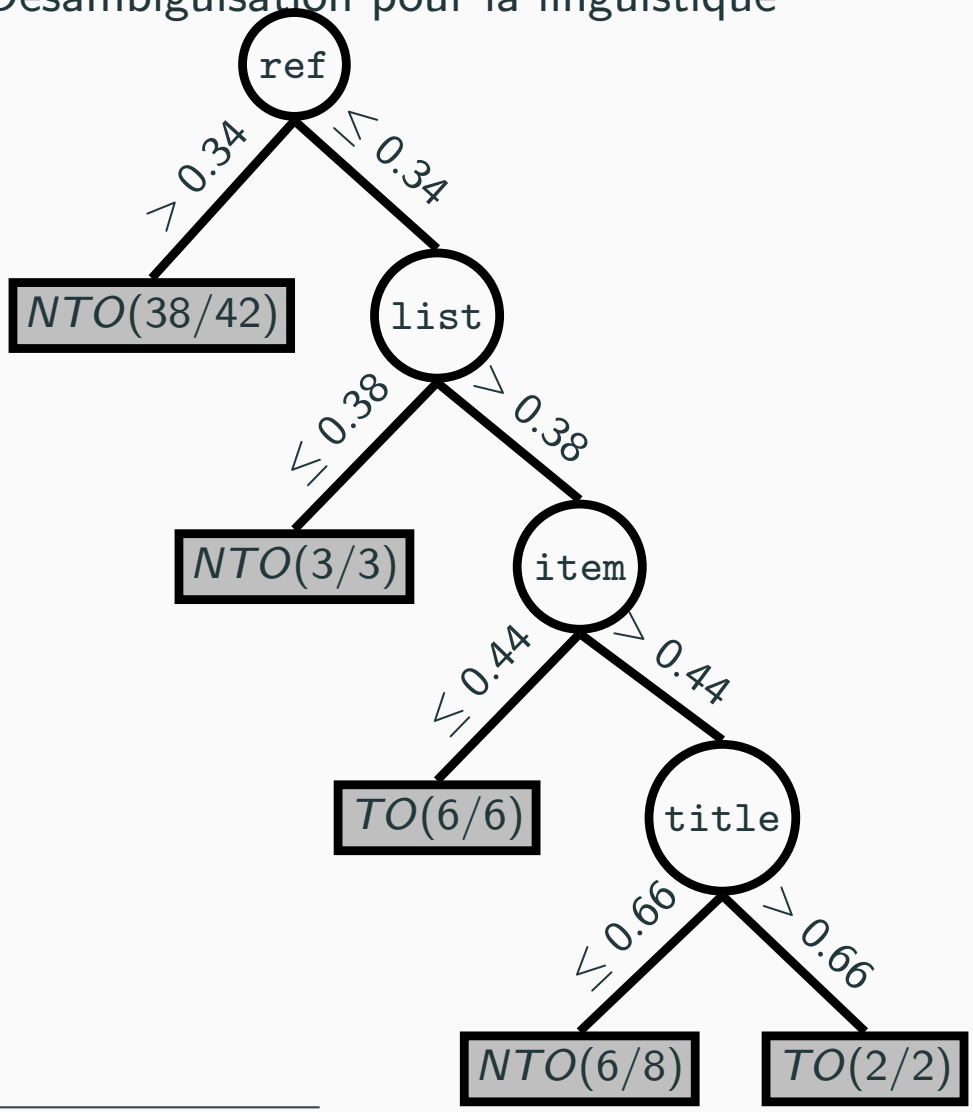
Tâche Désambiguïsation pour la linguistique [3]

3. [Daille et al., 2016]

Faut-il purger les XML ?

Données Corpus XML annoté terminologiquement

Tâche Désambiguïsation pour la linguistique [3]



3. [Daille et al., 2016]

Un plafond de verre pour les LLM ?

Données Corpus multilingue (el, en, fr, pl, ru, zh)

Tâche Classification et détection d'évènements[Mutuvi, 2022]

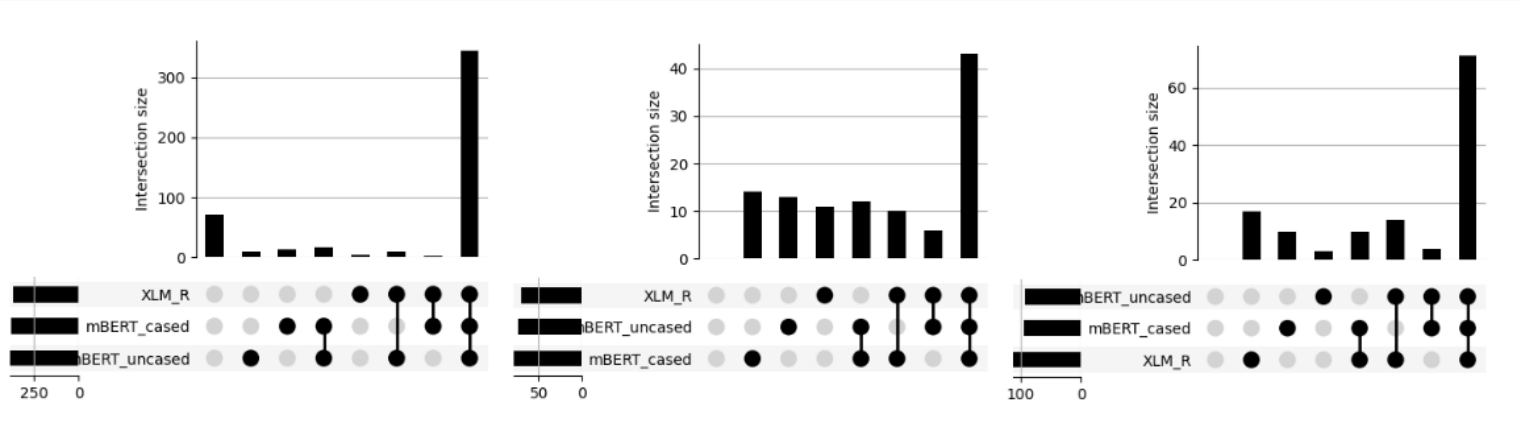


Figure 5 – Les modèles de langues sont faiblement complémentaires

Un plafond de verre pour les LLM ?

Données Corpus multilingue (el, en, fr, pl, ru, zh)

Tâche Classification et détection d'évènements[Mutuvi, 2022]

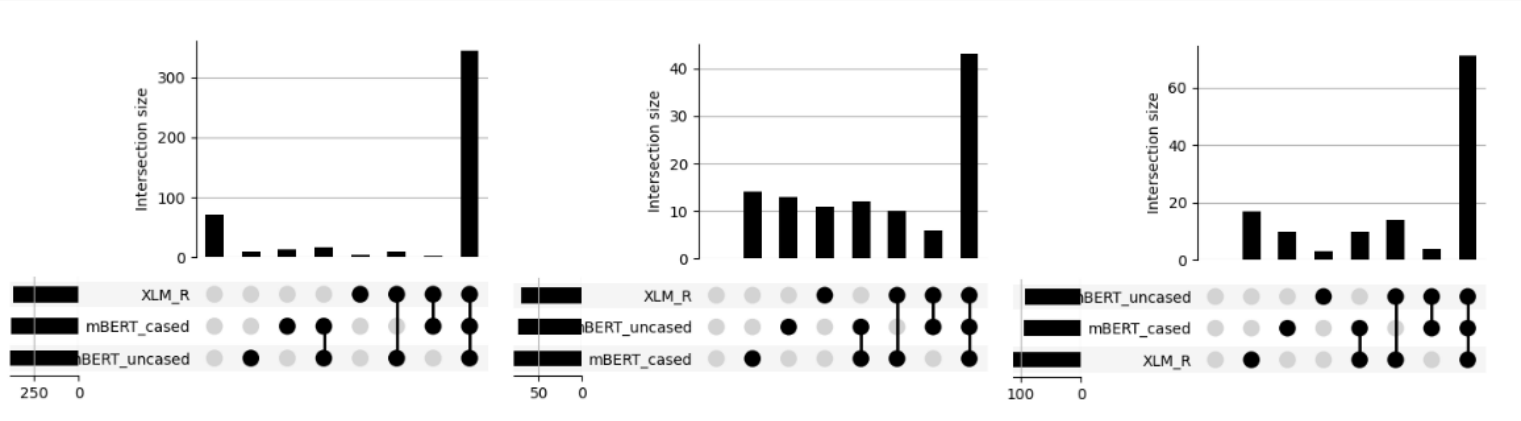


Figure 5 – Les modèles de langues sont faiblement complémentaires

Des heuristiques de "genre textuel" à la rescousse ? [Mutuvi et al., 2023]

Mais pourquoi genre et style ?

Genres textuels et caractéristiques stylistiques pour la classification

$\rightarrow$ un peu d'épistémologie/histoire du TAL

Flashback

L'équipe ISLAND du GREYC avec une épistémologie fondée sur trois principes :

- Non-compositionnelle
- Endogène
- Différentielle

Flashback

L'équipe ISLAND du GREYC avec une épistémologie fondée sur trois principes :

- Non-compositionnelle
- Endogène
- Différentielle

Conséquences :

- Irréductibilité du tout à la somme des parties [Lucas, 2009]
- Modélisation descendante plutôt qu'ascendante
- Propriétés Alingues [Vergne, 2003]
- Primauté du document sur la langue

Convenons toutefois que si le mot, ou mieux le morphème, est l'unité élémentaire, le texte est pour une linguistique évoluée l'unité minimale, et le corpus l'ensemble dans lequel cette unité prend son sens.[Rastier, 2005]

Non-compostionnel, endogène, différentiel

Il faut garder le jeu à l'échelle du joueur (Greimas ?)

Non-compostionnel, endogène, différentiel

Il faut garder le jeu à l'échelle du joueur (Greimas ?)

Dans la vision dominante en TAL, les formes sont réputées préexister et avoir des ≪ étiquettes ≫ indépendamment de la visée de l'application

. . . la question n'est pas posée en termes de trouver quelque chose de défini n'importe où, mais bien de trouver ce qu'il y a (indéfini) en un lieu défini.[Lucas, 2009]

Non-compostionnel, endogène, différentiel

Il faut garder le jeu à l'échelle du joueur (Greimas ?)

Dans la vision dominante en TAL, les formes sont réputées préexister et avoir des ≪ étiquettes ≫ indépendamment de la visée de l'application

. . . la question n'est pas posée en termes de trouver quelque chose de défini n'importe où, mais bien de trouver ce qu'il y a (indéfini) en un lieu défini.[Lucas, 2009]

. . . faire la part entre l'absence usuelle et l'absence marquée (nullax)[Pincemin, 2020]

Non-compostionnel, endogène, différentiel

Il faut garder le jeu à l'échelle du joueur (Greimas ?)

Dans la vision dominante en TAL, les formes sont réputées préexister et avoir des ≪ étiquettes ≫ indépendamment de la visée de l'application

. . . la question n'est pas posée en termes de trouver quelque chose de défini n'importe où, mais bien de trouver ce qu'il y a (indéfini) en un lieu défini.[Lucas, 2009]

. . . faire la part entre l'absence usuelle et l'absence marquée (nullax)[Pincemin, 2020]

les relations entre termes priment sur les termes eux-mêmes[Giguet, 2011]

Non-compostionnel, endogène, différentiel

Il faut garder le jeu à l'échelle du joueur (Greimas ?)

Dans la vision dominante en TAL, les formes sont réputées préexister et avoir des ≪ étiquettes ≫ indépendamment de la visée de l'application

. . . la question n'est pas posée en termes de trouver quelque chose de défini n'importe où, mais bien de trouver ce qu'il y a (indéfini) en un lieu défini.[Lucas, 2009]

. . . faire la part entre l'absence usuelle et l'absence marquée (nullax)[Pincemin, 2020]

les relations entre termes priment sur les termes eux-mêmes[Giguet, 2011]

Le global détermine le local[Lejeune, 2013]

Extraction terminologique et mots vides[Vergne, 2003]

(f = effectif, l = longueur)

	f	l	
1	649	3	*une*
2	32	8	*nouvelle*
3	1	10	*résolution*
4	3673	2	*de*
5	1500	2	*l'*
6	9	3	*ONU*

Extraction terminologique et mots vides[Vergne, 2003]

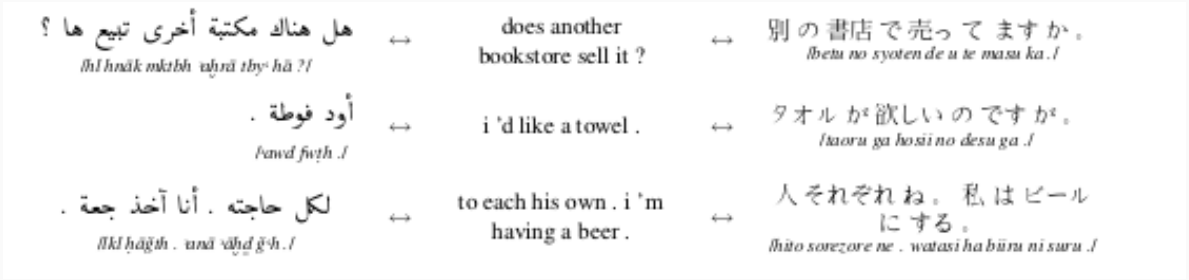
(f = effectif, l = longueur)

	f	l	
1	649	3	*une*
2	32	8	*nouvelle*
3	1	10	*résolution*
4	3673	2	*de*
5	1500	2	*l'*
6	9	3	*ONU*

0	v	4-27	like	ici, «like» est un mot vide
1	P	6-1	bamboo	
2	P	6-1	shoots	
3	v	5-11	after	
4	v	1-252	a	
5	P	6-1	spring	
6	P	4-1	rain	
0	v	3-33	But	
1	v	2-37	we	
2	P	4-27	like	ici, «like» est un mot plein
3	v	2-289	to	
4	P	3-4	buy	
5	-	5-16	those	occurrence indéterminée
6	P	10-8	businesses	
7	v	2-249	in	
8	v	1-252	a	
9	P	10-1	contrarian	
10	P	7-1	fashion	

	Manifestazioni	*per*	*la*	***pace***	*in*	***tutto***	*il*	***mondo***
longueurs	14	3	2	4	2	5	2	5
profils	long	court	court	long		long	court	long
effectifs	1	10	207	2	62	3	19	3
profils	rare	fréquent	fréquent	rare		rare	fréquent	rare
déductions		mot vide	mot vide				mot vide	

Alignement de n-grammes[Lardilleux and Lepage, 2008]



Alignement de n-grammes[Lardilleux and Lepage, 2008]

هل هناك مكتبة أخرى تبيع ها ؟	↔	does another bookstore sell it ?	↔	別の書店で売ってますか。
/hl hnāk mktbh ʾaḫrā tbyʿ hā ?/				/betu no syoten de u te masu ka ./
أود فوطة .	↔	i 'd like a towel .	↔	タオルが欲しいのですが。
/awd fwṭh ./				/taoru ga hosi i no desu ga ./
لكل حاجته . أنا آخذ جعة .	↔	to each his own . i 'm having a beer .	↔	人それぞれね。私はビールにする。
/lkl ḥāǧth . ʾanā ʾāḫḏ ǧʿh ./				/hito sorezore ne . watasi ha bīru ni suru ./

Arabic	English	Japanese	Freq.
	beer	ビール	202
بيرة	beer	ビール	35
بيرة	beer		14
البيرة	beer	ビール	8
البيرة	beer		8
جعة	beer	ビール	3
⋮	⋮	⋮	⋮

Arabic			Prob.
بيرة	/byrh/	'beer'	0.593
البيرة	/ālbyrh/	'a beer'	0.186
جعة	/ǧh/	'beer'	0.047
البيرة المحلية	/ālbyrh ālmḥlyh/	'local beer'	0.035

Japanese			Prob.
ビール	/bīru/	'beer'	0.970
国産ビール	/kokusan bīru/	'local beer'	0.019
缶	/kan/	'can (tin)'	0.004
しか	/sika/	'only'	0.003

Détection de Structure par le grain corpus[Giguet, 2011] (I)

8

1524 FLOETH AJNR: 22, September 2001

sentation of concomitant IMT-SPECT data, in VPC-mediated gene therapy. After local gene therapy, we noted a new type of dynamic, transient perifocal (flare) enhancement on MR images in certain patients. The IMT-SPECT investigations showed increased amino acid uptake in patients with enhancement from residual or relapsing tumor, but not in patients with flare. Thus IMT-SPECT may help to differentiate between tumorous and nontumorous flare enhancements in patients with enhancing tissue on MR images after gene therapy for glioblastoma.

The presence or absence of contrast enhancement within the suspected area of glioblastoma is thought to be a direct reflection of initial tumor size and the spread and development of tumor recurrence. The diagnostic value of contrast enhancement after surgical interventions, such as tumor resection, can become confusing. In these situations, enhancement may reflect residual tumor or it may reflect nonspecific enhancing phenomena such as inflammatory reactions or posttraumatic (postsurgical) tissue changes.

The so-called benign postsurgical enhancement is known to arise at the resection margins after surgical interventions such as tumor extirpations. Its characteristics include an absence of enhancement for the first 72 hours; development of a linear, thin enhancement at the edges of the resection cavity by the end of the first postoperative week; and persistence of the enhancement for a maximum of 3 months. It usually is not surrounded by substantial edema and does not cause mass shift. Potential underlying mechanisms include disruption of local blood-brain barrier, neovascularization, and luxury perfusion. Although residual tumor enhancement generally appears more nodular and irregular, postsurgical enhancement usually is linear. Knowledge of this phenomenon led to the recommendation that postoperative control images be obtained as early as possible after an operation, to distinguish between tumorous and nontumorous postsurgical enhancements (5). The reference standard for postoperative evaluation of residual tumor is thought to be an MR image within 48 hours after the procedure.

Benign postsurgical enhancement is not restricted to tumor surgery patients, however. It occurs after many different cerebral surgical procedures. In exceptional cases, benign enhancement can arise as early as 17 hours after nontumor surgery (such as temporal lobectomy for epilepsy) and it can appear to be both linear and nodular (11). This highlights that differentiation of contrast enhancement on postoperative MR images can be very difficult and that the presence of early postoperative enhancement alone does not always mean a diagnosis of residual neoplasm.

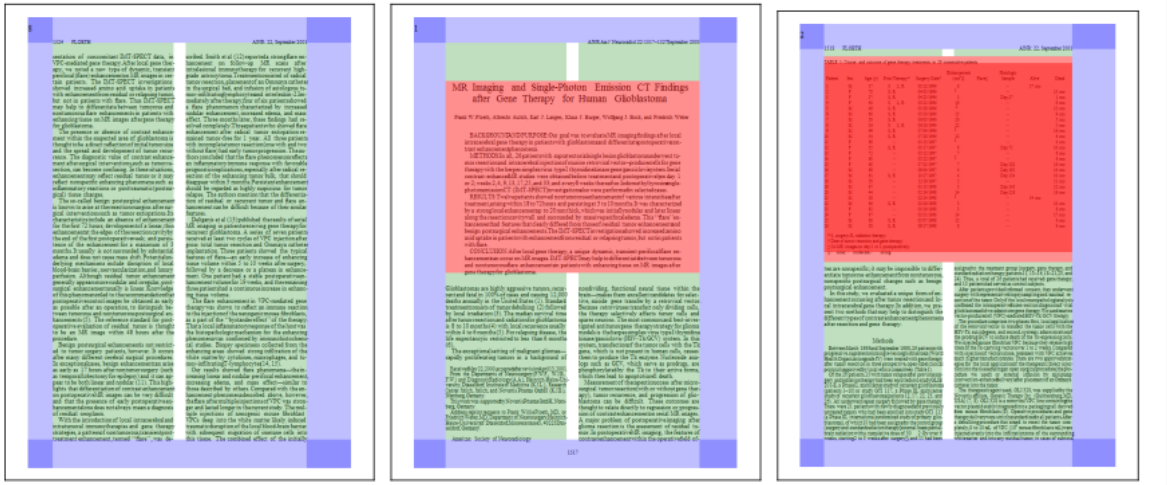
With the introduction of local intracerebral and intratumoral immunotherapies and gene therapy strategies, a pattern of nontumorous, transient posttreatment enhancement, termed "flare", was described. Smith et al (12) reported a strong flare enhancement on follow-up MR scans after intralesional immunotherapy for recurrent high-grade astrocytoma. Treatment consisted of radical tumor resection, placement of an Ommaya catheter in the surgical bed, and infusion of autologous tumor-infiltrating lymphocytes and interleukin-2. Immediately after therapy, four of six patients showed a flare phenomenon characterized by increased nodular enhancement, increased edema, and mass effect. Three months later, these findings had resolved completely. Three patients who showed flare enhancement after radical tumor extirpation remained tumor-free for 1 year. All three patients with incomplete tumor resection (one with and two without flare) had early tumor progression. The authors concluded that the flare phenomenon reflects an inflammatory immune response with favorable prognostic implications, especially after radical resection of the enhancing tumor bulk, that should disappear within 3 months. Persistent enhancement should be regarded as highly suspicious for tumor relapse. The authors mention that the differentiation of residual or recurrent tumor and flare enhancement can be difficult because of their similar features.

Deliganis et al (13) published the results of serial MR imaging in patients receiving gene therapy for recurrent glioblastoma. A series of seven patients received at least two cycles of VPC injection after gross total tumor resection and Ommaya catheter implantation. Three patients showed the typical features of flare—an early increase of enhancing tissue volume within 5 to 10 weeks after surgery, followed by a decrease or a plateau in enhancement. One patient had a stable postoperative enhancement volume for 19 weeks, and the remaining three patients had a continuous increase in enhancing tissue volume.

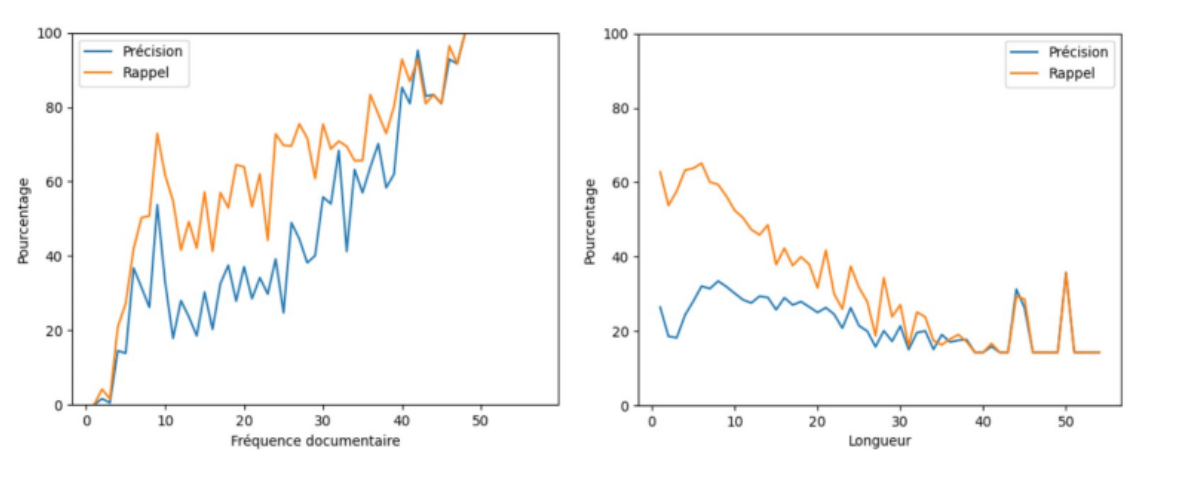
The flare enhancement in VPC-mediated gene therapy was shown to reflect an immune reaction to the injection of the xenogenic mouse fibroblasts, as a part of the "bystander effect" of the therapy. That a local inflammatory response of the host was the histopathologic mechanism for the enhancing phenomena was confirmed by immunohistochemical studies. Biopsy specimens collected from the enhancing areas showed strong infiltration of the white matter by cytokines, macrophages, and tumor-infiltrating T-lymphocytes (14, 15).

Our results showed flare phenomena—the increasing linear and nodular perifocal enhancement, increasing edema, and mass effect—similar to those described by others. Compared with the enhancement phenomena described above, however, the flare after multiple injections of VPC was stronger and lasted longer in the current study. The multiple injections of xenogenic mouse fibroblast VPCs deep into the white matter likely induced traumatic disruption of the local blood-brain barrier with subsequent migration of immune cells into this tissue. The combined effect of the initially

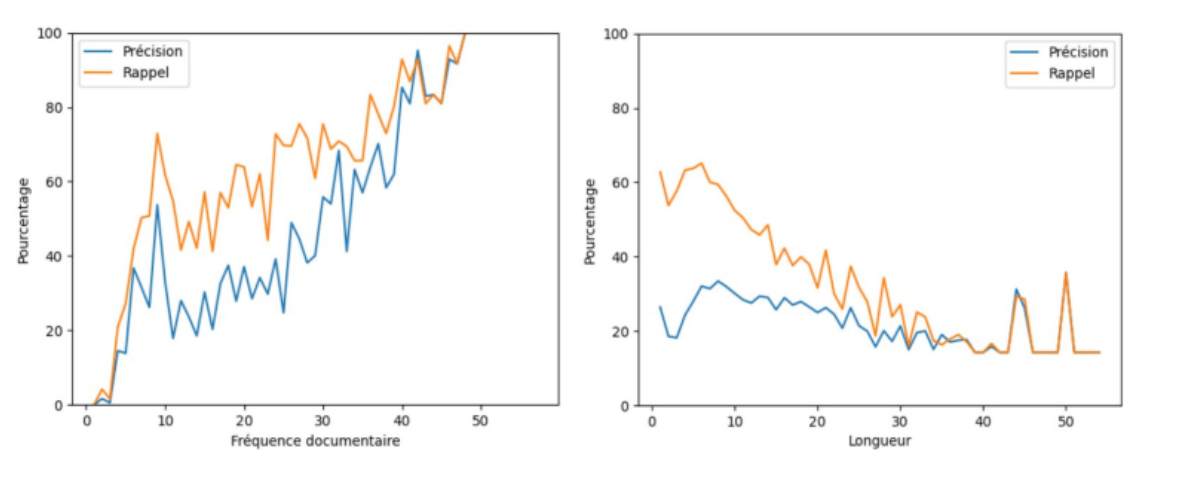
Détection de Structure par le grain corpus[Giguet, 2011] (II)

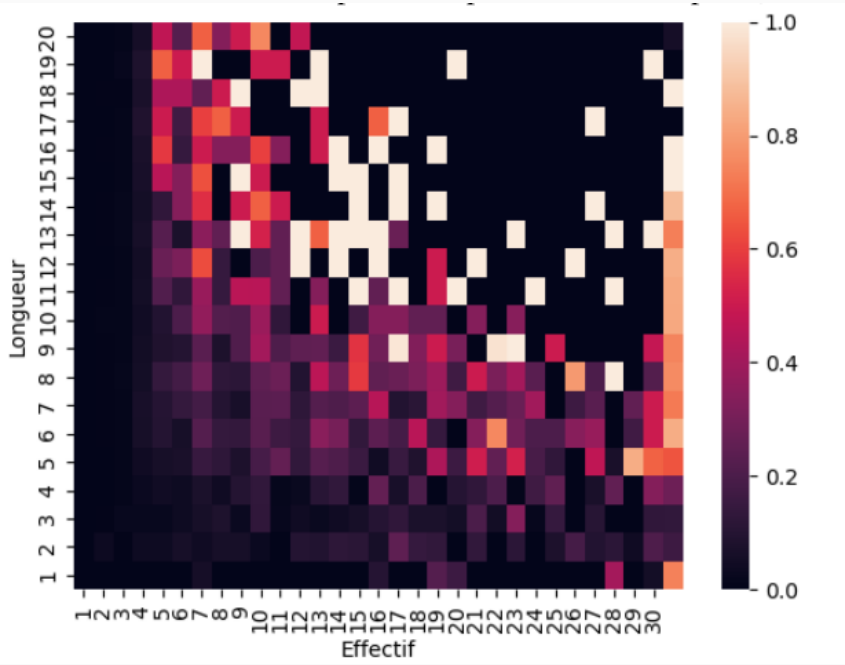


Longueur et fréquence, réminiscences (KoudoroParfait et al . 2025



Longueur et fréquence, réminiscences (KoudoroParfait et al . 2025





Réminiscience ?

Le traitement des langues s'est longtemps focalisé sur l'analyse d'écrits textbfpropres et lissés, débarrassés de toute scorie qui pourrait interférer avec les attendus des divers algorithmes. La stratégie dominante consiste à nettoyer le texte avant de l'analyser, à le **débarrasser de sa mise en forme**, à corriger ses fautes d'orthographe, à « traduire » ses sigles, ses abréviations, à le réduire à un français codifié, et plus généralement à une langue codifiée. L'attente forte sur la qualité du texte à analyser explique la dégradation des résultats lorsque l'on change de genre, et l'impossibilité de traiter les nouvelles formes de communication plus libres et non corrigibles.

Une conclusion sans conclusion

Faire du TAL sans étiquetage ?

Une conclusion sans conclusion

Faire du TAL sans étiquetage ?

Faire du TAL sans ML

Une conclusion sans conclusion

Faire du TAL sans étiquetage ?

Faire du TAL sans ML

Faire du TAL sans LLM ?

Une conclusion sans conclusion

Faire du TAL sans étiquetage ?

Faire du TAL sans ML

Faire du TAL sans LLM ?

Faire du TAL sans linguistique ? sans modèle ?

ça existe :

Références i

Baledent, A., Hiebel, N., and Lejeune, G. (2020).
Dating Ancient texts : an Approach for Noisy French Documents.
In Language Tech. for Historical and Ancient Languages, .

Daille, B., Jacquey, E., Lejeune, G., Melo, L. F., and Toussaint, Y. (2016).
Ambiguity Diagnosis for Terms in Digital Humanities.
In Language Resources and Evaluation Conference, Portorož, Slovenia.

Giguet, E. (2011).
De l'analyse syntaxique automatique à l'analyse automatique de discours dan

Accreditation to supervise research, Université de Caen Basse-Normandie.

Lardilleux, A. and Lepage, Y. (2008).
A truly multilingual, high coverage, accurate, yet simple, sub-sentential alignment method.
In Proceedings of AMTA 2008, pages 125–132, Waikiki, Honolulu, United States.

Références ii

Lejeune, G. (2013).
Veille épidémiologique multilingue : une approche parcimonieuse au grain caractère fondée sur le genre textuel.
PhD thesis, Université de Caen.

Lucas, N. (2009).
Modélisation différentielle du texte, de la linguistique aux algorithmes.
PhD thesis, Université de Caen.

Mutuvi, S. (2022).
Epidemic Event Extraction in Multilingual and Low-resource Settings.

Theses, La Rochelle Université.

Références iii

Mutuvi, S., Boros, E., Doucet, A., Jatowt, A., Lejeune, G., and Odeo, M. (2023).
Analyzing the impact of tokenization on multilingual epidemic surveillance in low-resource languages.
In Document Analysis and Recognition - ICDAR 2023, pages 17–32, Cham. Springer Nature Switzerland.

Nguyen, N. K., Boros, E., Lejeune, G., and Doucet, A. (2020).
Impact analysis of document digitization on event extraction.
In Natural Language for Artificial Intelligence (NL4AI), . -.

Pincemin, B. (2020).
La textométrie en question.
Le Français Moderne - Revue de linguistique Française, 88(1) :26–43.
numéro dirigé par Véronique Magri.

Rastier, F. (2005).
Enjeux épistémologiques de la linguistique de corpus.
La linguistique de corpus, (31-45).


Vergne, J. (2003).
Un outil d'extraction terminologique endogène et multilingue.
Actes de TALN, 2 :139–148.