



Usages variés de l'outil d'exploration de corpus TXM dans le contexte de l'allemand

Eva Schaeffer-Lacroix (Paris-Sorbonne/Espé de Paris) evalacroix@free.fr

Gilbert Magnus (Lille) gilbert.magnus@gmail.com

Thierry Grass (Strasbourg) tgrass@unistra.fr



Plan

- Vidéo d'illustration
- Arguments pour l'analyse textométrique
- Arguments pour l'utilisation de TXM
- Corpus
- Description linguistique
- Applications



Vidéo d'illustration





Arguments pour l'analyse textométrique

- Utiliser des outils de textométrie comme TXM permet de confronter des hypothèses linguistiques à une grande quantité d'occurrences et/ou d'en créer à partir des données observées.
- De tels outils permettent d'aborder le texte non seulement grâce à une lecture linéaire, mais aussi grâce à l'observation ciblée de certains de ses éléments et caractéristiques que l'on sélectionne en fonction des questions que l'on se pose.
- Quantifier des éléments d'un texte n'est pas une action récente (cf. concordances faites déjà au Moyen-Age pour la Bible). Elle est simplement devenue plus confortable grâce à la numérisation du texte et grâce aux outils d'interrogation.



Qu'est-ce que les analyses textométriques permettent de faire ?

- Décrire, au niveau local, des caractéristiques lexicales, syntaxiques, morphologiques, etc. d'un texte.
- Aller au-delà d'une description locale : étudier les relations entre éléments textuels (cooccurrences), les relations lexicales entre textes (spécificité) ; localiser des éléments textuels dans le corpus (progression).
- Travailler avec des annotations ≈ réfléchir aux catégories linguistiques.
- Appréhender la « résonance » de certains éléments linguistiques (André Salem).

(Voir aussi *Pincemin & Heiden 2008*).



Arguments pour l'utilisation de TXM

Logiciel...

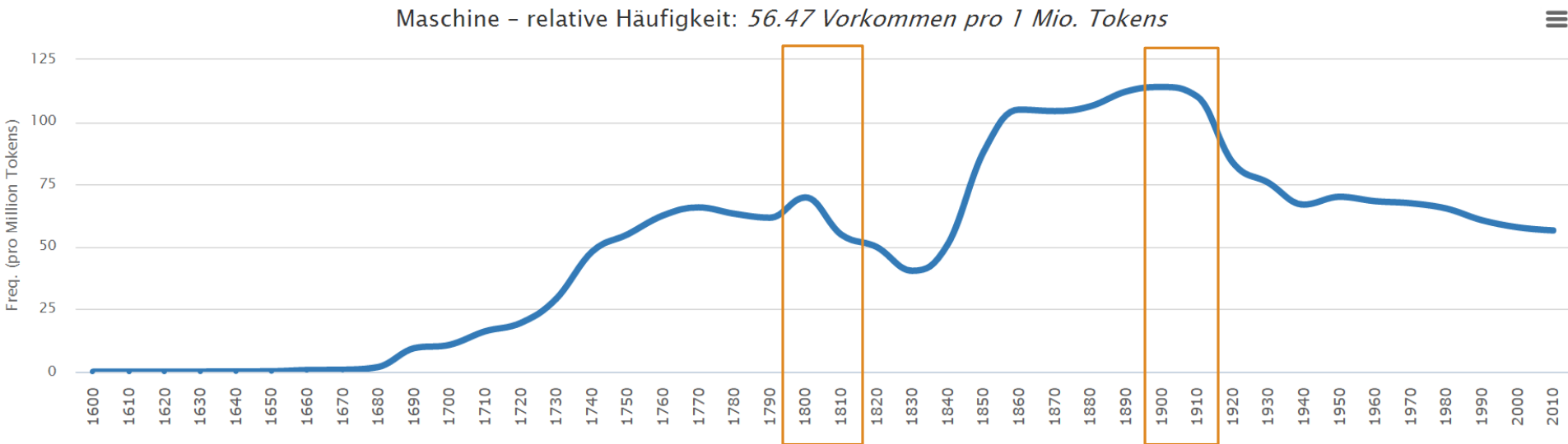
- Libre et gratuit
- En développement continu par une équipe de chercheurs/chercheuses de l'ENS de Lyon
- Permettant l'interopérabilité avec d'autres logiciels et formats d'édition courants
- Suffisamment facile à prendre en main tout en offrant une large gamme de fonctionnalités
- Aide rapide fournie aux usagers



Quelques fonctionnalités de TXM

Que peut-on faire ?	Précisions
Déposer, faire étiqueter, analyser ses propres corpus	Formats txt, csv, xml, etc.
Diviser le corpus en plusieurs parties	Fonctions « partition » et « sous-corpus »
Afficher la taille des parties d'une partition	Fonction « dimensions »
Calculer la fréquence d'un élément textuel	Fréquence absolue
Calculer la cooccurrence entre deux éléments textuels	La « compagnie » que se tiennent les mots
Calculer la spécificité du lexique d'une partition ou de sous-corpus	Le lexique « (a)typique »
Calculer la distance lexicale entre au moins quatre textes	Fonction « AFC » (Analyse Factorielle des Correspondances)
Localiser un ou plusieurs mots dans le corpus	Fonction « progression »
etc.	

« Maschine » dans *Digitales Wörterbuch der deutschen Sprache (DWDS)*





Corpus Science Fiction (SF)*

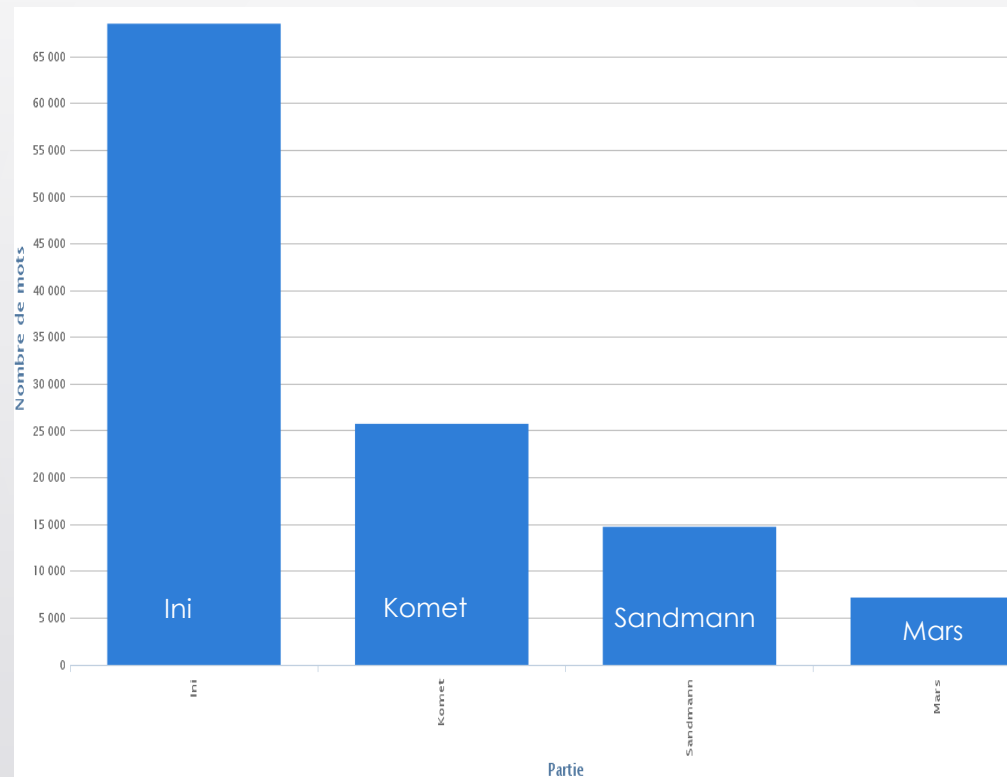
- VON VOß, Julius (1810). *Ini. Ein Roman aus dem 21. Jahrhundert.*
- HOFFMANN E.T.A. (1816-17). *Sandmann.*
- DOMINIK Hans (1908). *Die Reise zum Mars.*
- HEYMANN Robert (1909). *Der rote Komet.*

Nombre de mots du corpus entier : 116242

(max TXM = 10 millions, mais traitement plus difficile au-dessus de 3 millions)

*Source : *Project Gutenberg*

Taille des quatre parties du corpus



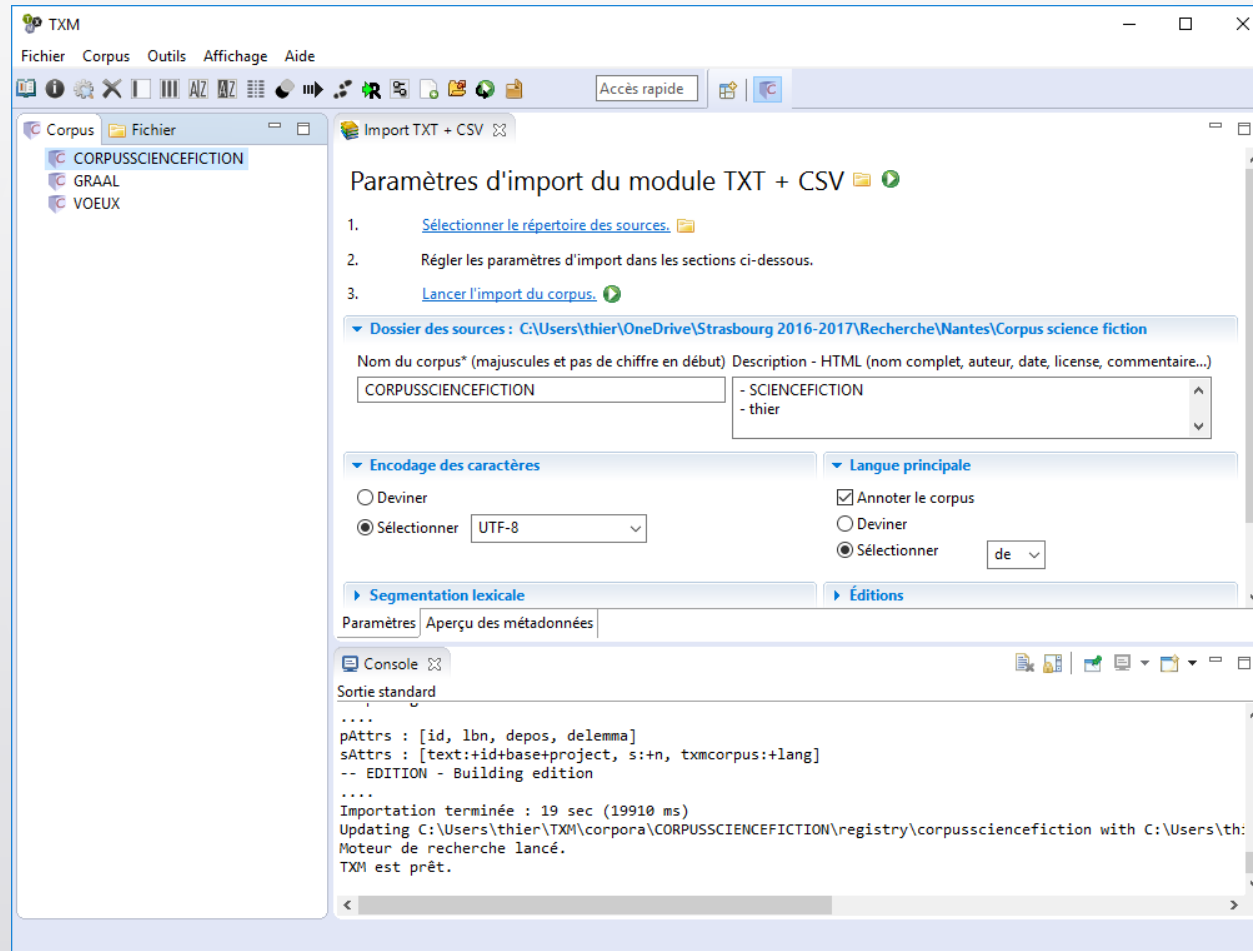
Calcul des spécificités

Unités	Fréquence T 116242	Komet t=25761	score	Mars t=7194	score	Ini t=68535	score	Sandmann t=14752	score
ADJA	6190	1234	-5,2	353	-1,3	3750	2,4	853	2,3
ADJD	3400	641	-6,1	206	-0,4	1948	-1,6	605	17,5
ADV	6898	1295	-12,2	488	3,0	4160	2,0	955	2,8
APPO	23	0	-2,5	1	-0,2	19	1,8	3	0,2
APPR	7381	1648	0,4	499	1,7	4431	1,6	803	-6,3
APPRART	996	138	-10,8	94	4,4	657	5,5	107	-1,5
APZR	18	8	1,5	3	1,0	4	-2,8	3	0,4
ART	10473	2307	-0,4	774	7,0	6451	8,4	941	-35,1
CARD	265	50	-1,0	64	20,5	130	-3,2	21	-2,0
FM	3	0	-0,3	1	0,8	2	0,2	0	-0,2
ITJ	33	1	-2,6	0	-0,9	8	-4,3	24	14,4
KOKOM	390	112	2,8	30	0,9	182	-6,2	66	2,0
KON	3428	782	0,7	233	1,1	1910	-4,3	503	3,6
KOUI	159	39	0,6	16	1,4	81	-1,6	23	0,5
KOUS	1700	423	2,4	99	-0,5	865	-11,1	313	11,1
NE	2677	908	44,6	85	-12,2	1248	-38,1	436	7,6
NN	20236	4112	-11,8	1419	7,1	12666	30,7	2039	-36,0

depos

Partie	Score
Komet	42.5
Mars	-12.5
Ini	-35.0
Sandmann	7.5

Import d'un corpus



Courte visite guidée de TXM

Ini. Ein Roman aus dem ein und zwanzigsten Jahrhundert
von
Julius v. Voß.
_Berlin, _ 1810.
Bei Karl Friedrich Amelang.

Vorrede.

Jean Paul sagt:» Friede mit der Zeit! sollte man öfter in sich hineinrufen.
Wie uns ein quälender Tag nicht in den Hoffnungen unsers Lebens irret, so
sollte uns ein leidendes Jahrhundert nicht die entziehen, womit wir uns die
weite Zukunft malen. «Wenn nun aber die Zeit gar unfriedlich ist, sollte da
nicht ein Blick in die Zukunft das bedrängte, oft zagende Herz trösten,
beleben, erheitern? Und eine bessere Zukunft naht so gewiß, als die
Vergangenheit von der Gegenwart übertroffen wird. Wenigstens gilt die
Behauptung, insofern wir, von der immer mehr entwickelten Kultur, das Heil
der Sterblichen erwarten. Was wir aber noch nicht sehen können, träumen,
ist ja wohl poetisch und religiös. Und
Sind's gleich nur Welten aus Ideen,
So baut man sie so herrlich als man will.
_Der Verfasser. _

Inventaire des mots

- Outils > Lexique : inventaire des mots rencontrés avec ordre de fréquence décroissante ou croissante.
- Plus haute fréquence (999999 – 101) : séparateurs, mots-outils, peu d'information sémantique.

word	Fréquence
,	10287
.	4832
die	2938
und	2577
der	2271
in	1566
zu	1359
den	1321
sich	1032
das	960
er	892
nicht	833
mit	826
von	816
sie	774
dem	756
auf	727
ein	661
!	573

Inventaire des mots

- Fréquence moyenne (100-10) :
vocabulaire central du corpus

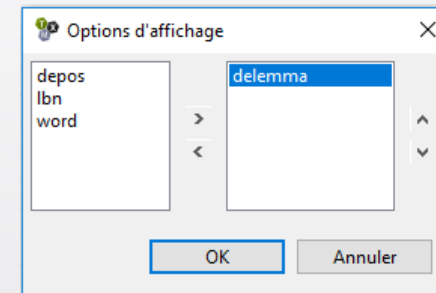
word	Fréquence
Monsieur	27
müssen	27
Nur	27
Raumschiff	27
Rom	27
standen	27
stehen	27
Tempel	27
wollten	27
zusammen	27
Bewegung	26

- Fréquence faible et hapax (9?-1)
vocabulaire périphérique, les
mots de fréquence 1 forment
environ la moitié du vocabulaire
du corpus

word	Fréquence
bewaffnet	1
bewaffnete	1
Bewaffnung	1
bewähren	1
bewahrt	1
bewährtesten	1
Bewallungen	1
Bewässerung	1
bewege	1
bewegende	1
bewegenden	1
bewegliche	1
beweglichen	1
Beweglichkeit	1
Bewegungstheorie	1

Lemmatisation vs. non lemmatisation

- Possible d'afficher les formes lemmatisées : avoir conscience que la lemmatisation est un filtre et que la textométrie prône la proximité par rapport au texte (certaines recherches seront plus fines sans lemmatisation), par contre intéressant pour les verbes pleins (de fréquence moyenne) les plus fréquents.
- Le vocabulaire permet de trouver des points d'entrée dans le corpus.



delemma	Fréquence
wenden	34
Bewegung	33
hören	33
voll	33
ach	32
Fenster	32
gelingen	32
Innere	32
leben	32
Müller	32
stark	32
Wesen	32
deutlich	31
dunkel	31
ergreifen	31
erklären	31

Concordances

- Une ligne de concordance se compose d'un pivot, représenté par un mot, un groupe de mots ou une expression régulière, et de son contexte droit et gauche. Ce dernier est affiché en fonction de paramètres divers, par exemple, le choix de la taille des contextes (en nombre de caractères) et leur tri.

Contexte gauche	Pivot	Contexte droit
durch die du Dr. Diabel sein Geheimnis	entreißen	könntest! «» Ich brauche keine Folter! Merke dir,
Luxus und Reichtum zu befriedigen. Peter Cornelius	entriß	dem Arbeiter, der neben ihm ging, die Fahne und stürzte
brüllend warf er sich auf den Freund,	entriß	ihm die Ohnmächtige und schlug ihm mit der geballten Faust ins Gesicht
es nicht, die Insel ihnen wieder zu	entreißen	. «Sizilien! o mein Sizilien! Ini, wo magst



Les expressions régulières

- Une expression régulière est une formalisation au moyen d'une chaîne de caractères correspondant à une syntaxe précise partagée par un certain nombre de mots ou groupes de mots.

Exemple : *erfahren, aufgefahren, fuhr*, etc.
correspondent à l'expression régulière
[delemma=".*fahren"]

Contexte gauche	Pivot	Contexte droit
Crofton.» Wir werden es wohl bald	erfahren	, «sagte Romulus Futurus nachdenklich. VI. Die Revolution hatte
Zeit die Mitte ihrer Reihen; Geschütze wurden	aufgefahren	, deren erste Schüsse allein etwa fünfzig der Feinde niederrissen. So
sogar in Flammen auf. In ein zweites	fuhr	der Strahl und tötete beinahe alle Bewohner, während zu gleicher Zeit
von Berlin war, ließ alle schweren Geschütze	zusammenfahren	und eröffnete eine furchtbare Kanonade gegen den Rauch, gegen die Wolken
und Tausende von großen Schiffen waren wie Sternschnuppen	niedergefahren	, brennende, leuchtende Klumpen, von denen sich Stoff-Fetzen und tote
ballten sich zusammen und die Nägel der Finger	fuhren	in sein Fleisch, seine Augen rollten. Er wurde so bleich

Composés de « Maschine » dans le corpus SF

Requête :  [delemma=".*(M|m)aschine.*"]

Seuils : Fmin : Fmax : Vmax : Résultats par page :

|< < 1 -8 / 8 > >|

delemma	Fréquence T=118332	Komet t=25759	Mars t=7194	Ini t=68524	Sandmann t=16855
Maschine	14	2	1	10	1
Maschinenwerk	2	0	0	2	0
Hebemaschinen	1	0	0	1	0
Kleidungsmaschinen	1	0	0	1	0
Lademaschinen	1	0	0	1	0
Maschinenruderwerken	1	0	0	1	0
Maschinenschrift	1	1	0	0	0
Pflanzmaschinen	1	0	0	1	0

Étude des cooccurents de « Mensch » dans le corpus SF

Requête :

Propriétés des cooccurents : delemma Seuils : Fmin ≥ Cmin ≥ Indice ≥

Contexte : ☒ forme ☐ structure ☒ Contexte gauche actif ☒ Contexte droit actif ☐ inclure la structure contenant le pivot

de - à - et de à

Cooccurrent	Fréquence	Cofréquence	Indice	Distance moyenne
die	9712	222	4	4,1
Hundert	16	4	3	3,0
Todesgeschrei	2	2	3	1,5
antworten	24	4	3	6,8
Geschrei	3	2	3	2,5
schreiend	3	2	3	1,0
prallen	3	2	3	8,0
Erde Erden	4	2	2	1,0
;	192	10	2	5,3
keine	192	10	2	2,6
Feuerkreis	5	2	2	7,0
poetisch	5	2	2	4,0
ehe	18	3	2	6,0
ein	84	6	2	4,7
Wille	19	3	2	8,0
Haus	64	5	2	6,4

Les noms les plus fréquents(> 80) dans le corpus SF

Requête :

Seuils : Fmin : Fmax : Vmax : Résultats par page :

|< < 1 -13 / 13 > >|

delemma	Fréquence T=116242	Komet t=25761	Mars t=7194	Ini t=68535	Sandmann t=14752
Auge	179	76	4	41	58
Zeit	169	42	14	110	3
Jahr	109	9	19	74	7
Tag	107	23	10	60	14
Hand	98	38	3	33	24
Mann	95	25	0	63	7
Frau	93	83	0	7	3
Mensch	93	42	1	37	13
		18	2	48	24
		12	0	68	12
		36	22	25	2
		0	0	46	35
		40	15	16	9

	Fréquence relative			
Lemme	Komet	Mars	Ini	Sandmann
Auge	2949,56	556	598,19	3931,24
Zeit	1630,02	1946	1604,9	203,34

Thématiques du corpus

- Les principaux champs lexicaux définis à partir des noms les plus fréquents permettent de déterminer les thématiques principales du corpus.

Champs lexicaux	Occurrences dans le corpus
Le temps	<i>Zeit, Jahr, Tag, Augenblick</i>
L'être humain	<i>Auge, Hand, Mann, Frau, Mensch, Leben, Liebe, Vater</i>

Fréquence de [delemma=".*(Z | z)eit.*"] dans SF

word	Fréquence T=116228	komet t=25759	mars t=7194	ini t=68524	sandmann t=14751
Zeit	136	39	11	83	3
Zeiten	33	3	3	27	0
Vorzeit	12	0	0	12	0
zeitig	5	0	0	5	0
Gleichzeitig	4	4	0	0	0
Zeitalter	4	0	0	4	0
zeither	4	0	0	4	0
Zeitpunkt	4	0	3	1	0
gleichzeitig	3	1	2	0	0
Zeitalters	3	0	0	3	0
Zeitraume	3	0	0	3	0
Zeitungen	3	3	0	0	0
Dienstzeit	2	0	0	2	0
Jugendzeit	2	0	0	0	2
Lehrzeit	2	0	0	2	0
Zeitgeist	2	0	0	2	0
Zeitgenossen	2	0	0	2	0
zeitiger	2	0	0	2	0
Zeitraum	2	0	1	1	0

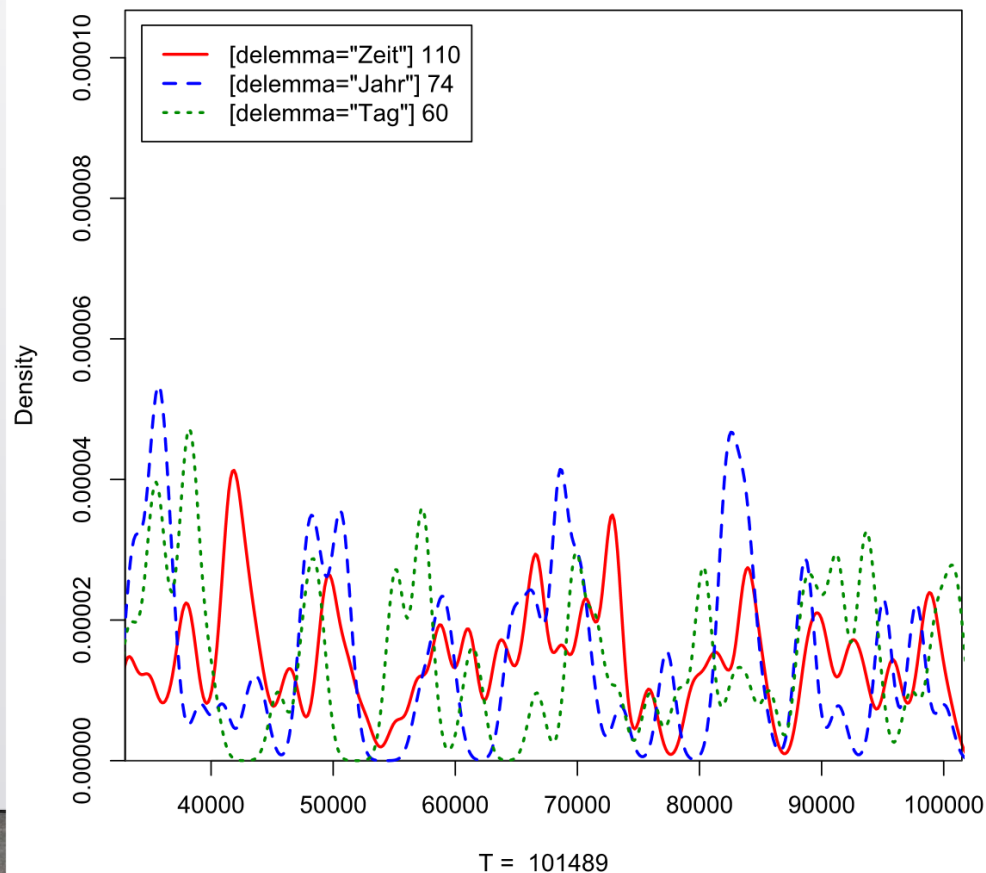
Lignes de concordances de [delemma=".*(Z | z)eit.*"] dans *Ini*

Contexte gauche	Pivot	Contexte droit
:» Friede mit der	Zeit	! sollte man öfter in sich hineinrufen. Wie
«Wenn nun aber die	Zeit	gar unfriedlich ist, sollte da nicht ein Blick
flammender Lebenskraft verbundenen wilden Neigungen	zeitig	zu entwaffnen, wenn nicht folgender Umstand hinzugetreten wäre
Neben Gelino wohnte seit einiger	Zeit	die edle Athania, Wittwe des afrikanischen Helden Medon
, und die Formen dieses	Zeitalters	zu betrachten. Die Schädelkunde, am Ende des
Weisheit_, der in früheren	Zeiten	nicht lebend anzutreffen war, den auch die Künstler
dir zwei bis drei Jahre	Zeit	, hast du dann diese Schönheit dir anezogen,
soll dich in dem langen	Zeitraume	nicht erblicken ? » Die erste Prüfung ! Auch
aber sie galten meistens Denkmäler	vorzeitlicher	Irrthümer. Die wenigen, welche in den Tagen
Mund. -- Von der	Zeit	an sahen sich die Liebenden öfter. Viel inniger
, da Guido in diesem	Zeitraume	fast seine Wohnung nicht mied. Auch Athania,
des Eilandes, und kurze	Zeit	darnach lag es in seiner ganzen Gestalt erkennbar unter
Inis Erzieherin, es ist	Zeit	, daß du an die Rückkehr denkst, Guido
. Man hatte um diese	Zeit	die Mittel, sich aus der Ferne zu unterhalten

Von Voß, Julius (1810). *Ini. Ein Roman aus dem 21. Jahrhundert.*

Progression – Zeit, Jahr, Tag (Ini)

Progression de [delemma="Zeit"], [delemma="Jahr"], [delemma="Tag"] dans Ini (densité)





Faire utiliser TXM par des étudiant·e·s

- Dans l'idéal, utilisation de la version *TXM* en ligne qui doit être implémentée sur un serveur de l'université
- En attendant, installation de TXM dans les ordinateurs des étudiant·e·s
- Mise à disposition de fichiers prêts à l'emploi contenant un corpus et ses métadonnées



Application dans le domaine de l'apprentissage des langues

- Réfléchir sur la langue
 - Observer les annotations proposées par *TreeTagger* (Schmidt, 1993-1996) et les soumettre à discussion.
 - Découvrir des caractéristiques morphosyntaxiques (quelle préposition ? quel ordre des mots ?)
- Réviser sa production écrite
 - Ponctuation
 - Collocations
 - Quasi-synonymie



Application en linguistique textuelle

Découvrir les caractéristiques de genres textuels

- Structure interne : dialogues *versus* didascalies (script de film)
- Lexique spécifique
- Récurrence de certaines unités lexicales
- Collocations
- Champs sémantiques développés
- Présence ou non de connecteurs et/ou de mots du discours.
Lesquels ?
- La cohésion est-elle assurée grâce aux relations anaphoriques et cataphoriques ? Construction progressive du sens ?
- Temps et modes utilisés



Pour finir

- Approche demandant un certain effort pour la prise en main de l'outil, mais offrant l'occasion de réaliser des analyses linguistiques variées
- Projet de proposer des ateliers TXM pour germanistes
- Analyse textométrique de traductions ([site Thierry Grass](#))
- [Club Corpus](#), laboratoire CeLiSo de Paris-Sorbonne
- [Vidéos atelier d'initiation à TXM](#) (Bénédicte Pincemin, 2012)



Références

- Brunet, Etienne (2013). "[What Do Statistics Tell Us ?](#)". In Hockey, S., Ide, N., Lancashire, I. (dir.). *Research in Humanities Computing*, vol. 1. Clarendon Press : Oxford, 1991. pp. 70-92.
- Heiden, S., Magué, J-P., Pincemin, B. (2010). "[TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement](#)". In Bolasco, S. (dir.). *Proc. of 10th International Conference on the Statistical Analysis of Textual Data. JADT 2010*. Vol. 2. pp. 1021-1032. Rome : Edizioni Universitarie di Lettere Economia Diritto.
- Klein, Wolfgang & Geyken, Alexander (2010). "Das Digitale Wörterbuch der Deutschen Sprache (DWDS)". In Heid, Ulrich, Schierholz, Stefan, Schweickard, Wolfgang, Wiegand, Herbert Ernst, Gouws, Rufus H. & Wolski, Werner (dir.). *Lexikographica*. pp. 79-93.
- Pincemin, B. (2012). [Tutoriel pour la prise en main de TXM](#).
- Pincemin, B. & Heiden, S. (2008). "Qu'est-ce que la textométrie ? Présentation", *Site du projet Textométrie*. <http://textometrie.ens-lyon.fr/spip.php?rubrique80>
- Schmid, H. (1993-1996). *Treetagger*. Outil d'annotation. <http://www.ims.uni-stuttgart.de/projekte/corplex/Treetagger/>



Publications de l'équipe TXM

- Heiden Serge (2010) « [The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme](#) », in 24th Pacific Asia Conference on Language, Information and Computation, Sendai, Japon. Institute for Digital Enhancement of Cognitive Development, Waseda University, pp. 389-398.
- Heiden Serge, Magué Jean-Philippe et Pincemin Bénédicte (2010). « [TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement](#) ». JADT 2010 : 10th International Conference on the Statistical Analysis of Textual Data, Rome, Italie. Edizioni Universitarie di Lettere Economia Diritto, 2 (3), pp.1021-1032, 2010. <halshs-00549779>
- Lay Marie-Hélène et Pincemin Bénédicte (2010). « [Pour une exploration humaniste des textes : AnaLog](#) ». JADT 2010 : 10th International Conference on Statistical Analysis of Textual Data, Rome, Italie, pp. 1045-1056.
- Pincemin Bénédicte (2010). « [Fonctionnalités textométriques : synthèse et innovations](#) ». Powerpoint présenté au séminaire du chantier Textopol, Université Paris 12 Créteil.
- Pincemin, Bénédicte (2012). Sémantique interprétative et textométrie. *Texte ! Textes et Cultures*, Equipe Sémantique des textes, 2012, XVII (3), pp.1-21. <halshs-00981180>
- Pincemin Bénédicte, Guillot Céline, Heiden Serge, Lavrentiev Alexei et Marchello-Nizia Christiane (2008). « [Usages linguistiques de la textométrie. Analyse qualitative de la consultation de la Base de Français Médiéval via le logiciel Weblex](#) ». In Mathieu VALETTE (dir.). *Textes, documents numériques, corpus. Pour une science des textes instrumentée - Syntaxe & Sémantique* 9, pp. 87-110.
- Pincemin Bénédicte, Heiden Serge, Lay Marie-Hélène, Leblanc Jean-Marc et Viprey Jean-Marie (2010). « Fonctions textométriques : Proposition de typologie selon un point de vue utilisateur ». JADT 2010 - Statistical Analysis of Textual Data - Proceedings of 10th International Conference, Rome : Edizioni Universitarie di Lettere Economia Diritto. pp.341-353. <halshs-00856446>