Quelques propositions d'expériences liées au langage

Éric de la Clergerie
<Eric.De_La_Clergerie@inria.fr>

http://alpage.inria.fr
INRIA Paris-Rocquencourt / Univ. Paris Diderot

Rencontre ISN
Rocquencourt – 8 Avril 2014

La dernière fois

Paul, je t'ai dit que François Flore est sorti faché de chez son banquier car celui-ci lui avait ex abrupto refusé son prêt pour sa future maison ?

Pragmatique: contexte & connaissances
référants: celui-ci=banquier, lui=son=sa=François, t'=Paul
structures argumentatives: refus explique faché
scénarios, implicites

Sémantique: sens des énoncés et des mots
structure prédicatives, rôle des actants (agent, patient, ...)
`refuser(agent=celui-ci,patient=lui,theme=prêt)`

Syntaxe: structure des phrases et relations entre mots
fonctions syntaxiques (sujet, objet, ...) : **celui-ci**=sujet, **prêt**=objet, **lui**=obj indirect de **refusé**

Morphologie: les mots et leur structure (**lubéronisation**)
découpage du texte en mots, catégories syntaxiques:
`celui/pro -ci/adj lui/cld avait/aux ex_abrupto/adv ...`
flexion (conjugaison) : **avait**=`avoir+3s+Ind+Imparfait`
entitées nommées (personnes, lieux, ...) : (François Flore) `PERSON_m`

Aujourd'hui

Objectif: Essayer de cerner la structure des langues

- avec des méthodes simples mais finalement puissantes
- pour étudier richesse et prédictibilité

Méthodes:

- caractères, séquences de caractères (n-gramme), mots
- fréquences
- probabilités
- modèles de langue

Linguistique quantitative, linguistique dirigée par les données, linguistique de corpus.

Expériences

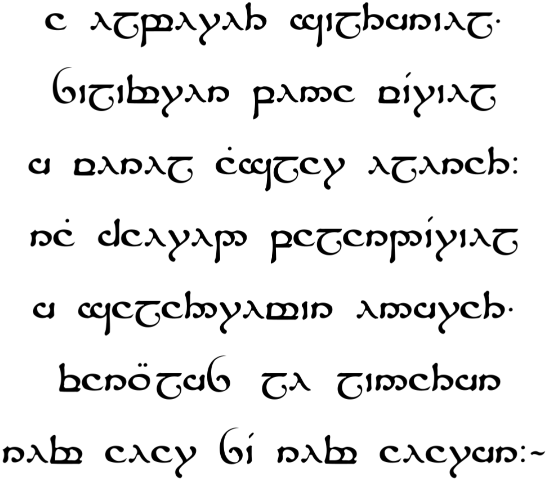
Utilisant de documents récupérés sur le Projet Gutemberg
`http://www.gutenberg.org`

- en français: Jules Vernes, Proust, Maurice Leblanc, Gaston Leroux, Stendhal ($\sim$ 1Mots)
- en anglais: les oeuvres de Shakespeare ($\sim$ 1Mmots)

Quelques scripts écrits en Perl

- disponibles sur demande
- langages alternatifs (calcul scientifique & statistiques) : Python (numpy), R, Octave, . . .

Plan



- Message A

  Les blaireaux viennent de gagner une bataille décisive au Royaume-Uni.

- Message B

  uyf pven-yexo anyccycb gy 3e3cy- xcy pebenvvy gs'nfnay ex UdlexqyiAcn.

- Message C

  éev -dfvonèné axeé3o't -t èfjvmv ec3 galqjvfu bmlpspcb è3 UpcuèuAb3ix.

- Message D

  Aq'sRv AUxUplRv-URèlquyci q3dppgciyx-Uxsln AUmp lqplbbRv3fRv dlgUyx iAf-iqAqbbRvpl-U 3p3fApstjsstgU3p lqyx -lstgU'glq-Ufm3pyxx-dp.

Entropie

Les langues naturelles exhibent un mélange typique:

- de redondance
  mots fonctionnels (articles, prépositions, conjonctions, . . .) et de mots très fréquents
- de diversité (richesse du vocabulaire et des constructions)
- + des distributions sur les longueurs des mots
  les mots fréquents sont souvent courts

$\Longrightarrow$ impact sur l'entropie des messages

Base: Mesure de l'entropie de l'anglais par Shannon
Prediction and Entropy of Printed English (1950)

Calcul de l'entropie

Point de départ: dans quelle mesure peut-on prédire le caractère c_{n+1} prolongeant une séquence $c_1 \cdots c_n$

- complètement aléatoire *fdabRr pne-ba-RècU*
- complètement prédictible *abababababab*
- en partie prédictible *je me demande ce qu*

Plus formellement:

$$H = \lim_{n \to \infty} H_n$$

avec

$$H_{n+1} = -\Sigma_{c_1 \cdots c_n c_{n+1}} p(c_1 \cdots c_n c_{n+1}) \log_2 p(c_{n+1} | c_1 \cdots c_n)$$

Cas limites:

- $H_0 = \log_2 |\text{alphabet}|$ (distribution équiprobable)
- $H_1 = -\Sigma_c p(c) \log_2 p(c)$

Calcul de l'entropie en pratique

Les entropies H_n sont calculées sur de grands corpus en prenant:

$$p(c_1 \cdots c_n) = \frac{\#(c_1 \cdots c_n)}{\#(\text{séquences de taille } n)}$$

Problèmes:

- le nombre de séquences croit très vite avec n
  $\Longrightarrow$ coût en temps et place
- Jamais assez de données pour observer suffisamment d'occurrences de $c_1 \cdots c_n$ pour n un peu grand
  $\Longrightarrow$ techniques de lissage (non utilisées aujourd'hui)

Note Google distribue des n-grammes calculés sur des corpus gigantesques pour diverses langues
`https://books.google.com/ngrams`

Résultats

> cat *.l1.fr | perl ./entropy.pl 4

H_n	en	fr	B	C	D	rand(a,b)	a^*
0	6.53	7.17	7.16	7.16	7.17	1.00	0.00
1	4.73	4.47	4.47	6.59	6.61	1.00	0.00
2	3.60	3.48	3.48	6.48	4.36	1.00	0.00
3	2.82	2.76	2.76	6.08	3.81	1.00	0.00
4	2.24	2.22	2.22	3.01	3.57	0.99	0.00
5	1.87	1.82	1.82			0.99	0.00

Pour l'anglais (27 caractères), Shannon trouve $H_3 = 3.3$
et postule une entropie H entre 1 et 2.
utilisation d'un jeu de déduction (type pendu)

Pour H_0, codage des caractère sur 7 ou 8 bits. Moins de bits nécessaires en considérant des séquences $\Longrightarrow$ compression.

Pour aller plus loin

L'entropie n'est qu'un premier pas pour déterminer le statut d'un message.

D'autres indices possibles

- diversité des mots
- fréquence d'émergence de nouveaux mots
- lien entre fréquence et longueur des mots
- taux d'utilisation de l'espace potentiel des mots
- ...

Loi de Zipf

Loi en puissance très présente dans les données linguistiques, traduisant une décroissance exponentielle de la fréquence f en fonction du rang r:

$$f_r \sim \frac{1}{r^\alpha} \quad \alpha > 1$$

- quelques mots/structures sont beaucoup utilisés; énormément de mots/structures sont très peu utilisés
- traduit à la fois une prime à la réutilisation et une tendance à la créativité

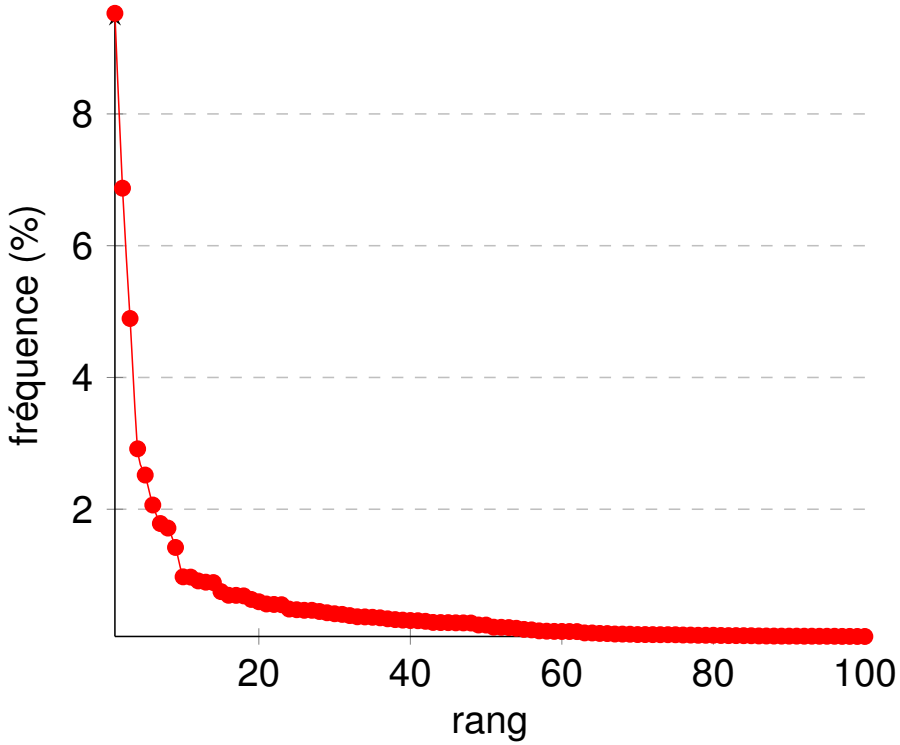
Note: relation similaire sur les longueurs des mots

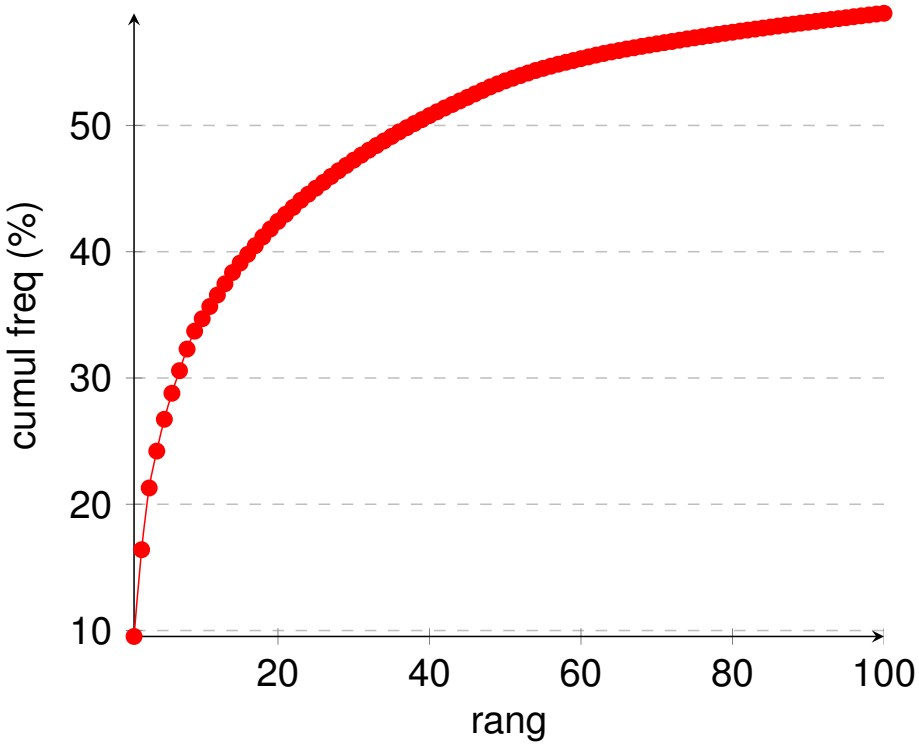
$$l = 1 + \frac{a}{f^b}$$

les mots fréquents sont courts en moyenne

Distribution de lemmes

Distribution des mots (lemmes) dans un corpus de 500 millions de mots, avec 3 234 274 lemmes distincts dont 71 348 hors noms propres:

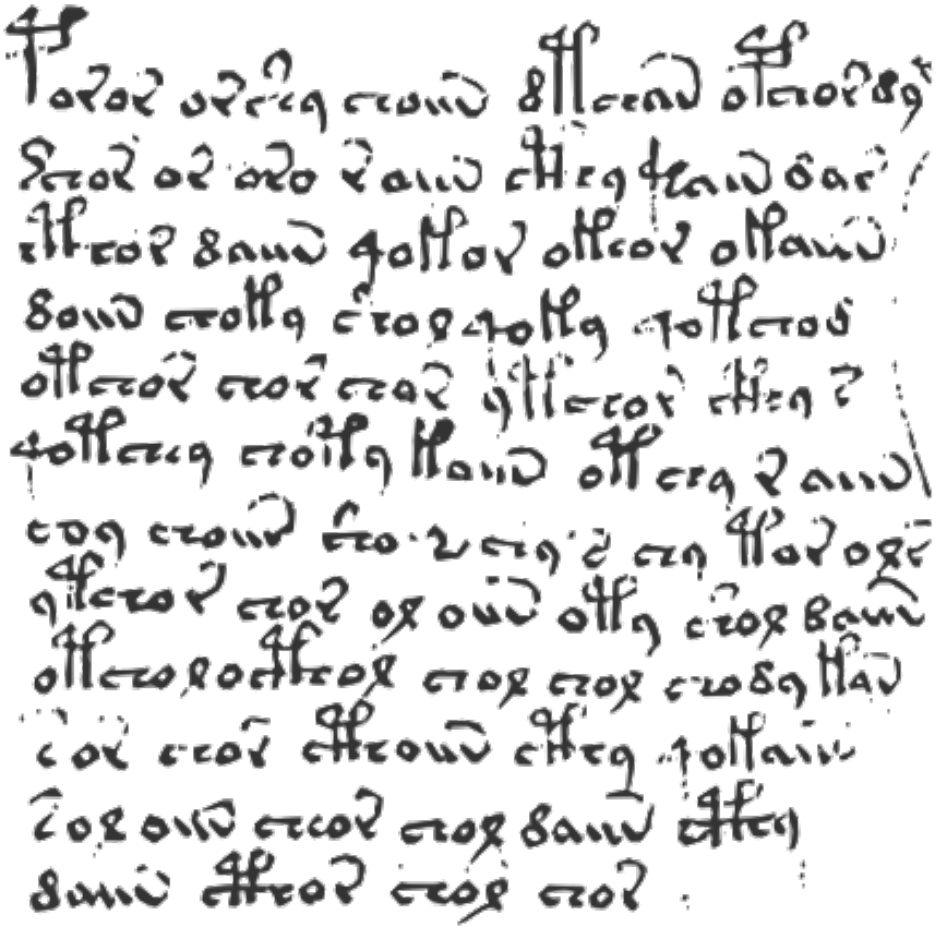




Les mots les plus fréquents: **le**, **de**, “,”, “.”, **à**, **un**, **et**, `cln`, “:”, **en**, **être**/v, . . .
80% des occurrences couvertes avec ∼1500 lemmes et 90% avec 6000 mots

Le manuscript de Voynich

Livre de 234 pages écrit entre 1450 et 1520, avec illustrations, mais auteur inconnu et contenu non déchiffré. Mais respect des critères d'une langue.
`http://fr.wikipedia.org/wiki/Manuscrit_de_Voynich`



Plan

Une tâche facile

Outils:

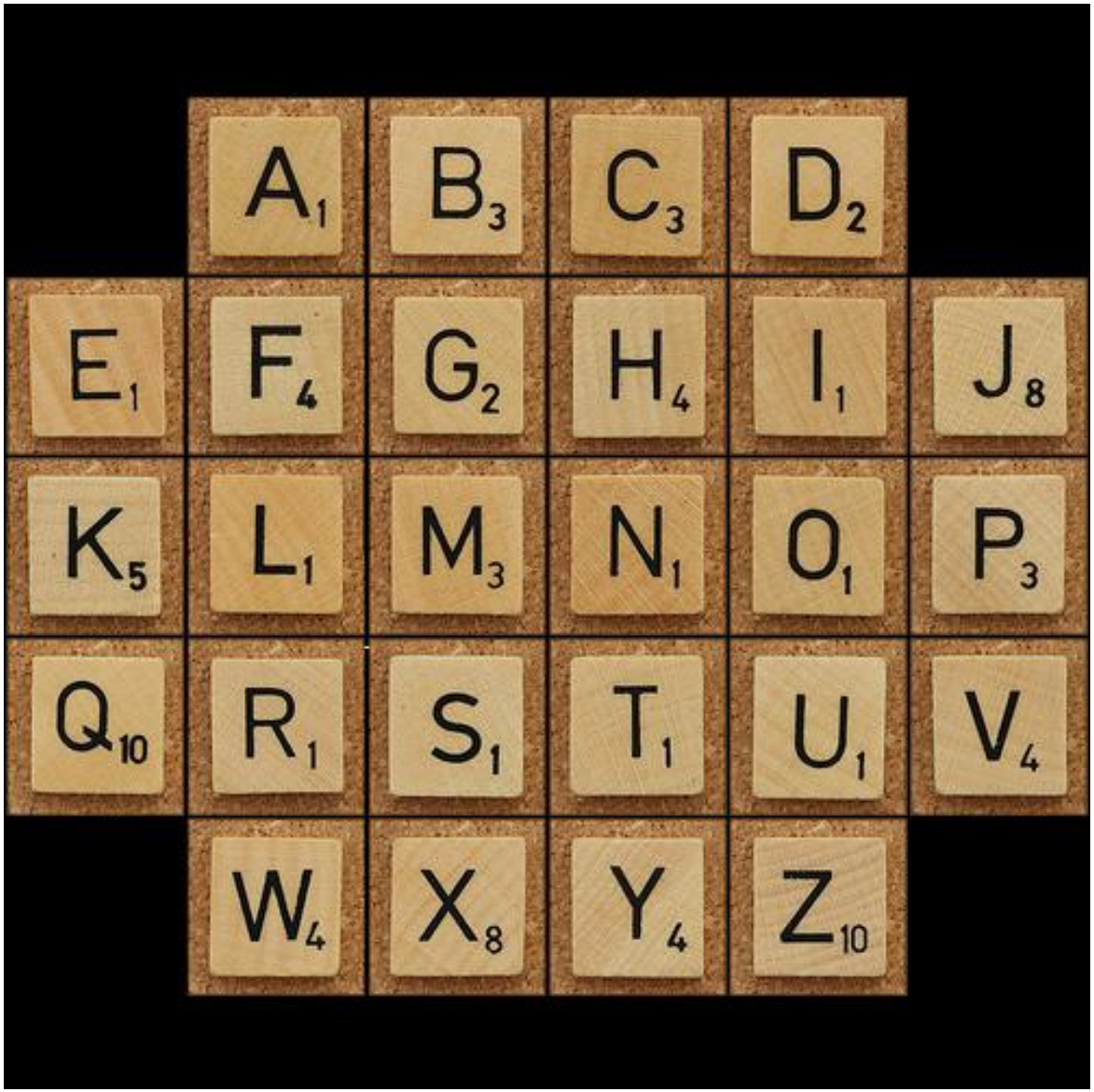
- en ligne: `http://whatlanguageisthis.com/`
- libres: MGUESSER `http://www.mnogosearch.org/guesser/`

```
> echo "Beware the Jubjub bird , and shun The frumious
    Bandersnatch" | ./mguesser -d maps/ -n3
0.6202442646 en iso-8859-1
0.6046028733 de latin1
0.5912522078 fr utf8
```

```
> echo "Il était grilheure ; les slictueux toves Gyraient sur l'
    alloinde et vriblaient" | ./mguesser -d maps/ -n3 -l l1
0.6878187060 fr utf8
0.6851934791 fr latin1
0.6823609471 fr iso-8859-1
```

```
> echo "Nakita kitá sa tindahan kahapon" | ./mguesser -d maps -n3
0.5999047756 tl ascii
0.5547670126 tl ascii
0.5282356739 fi latin1
```

Statistiques sur les caractères

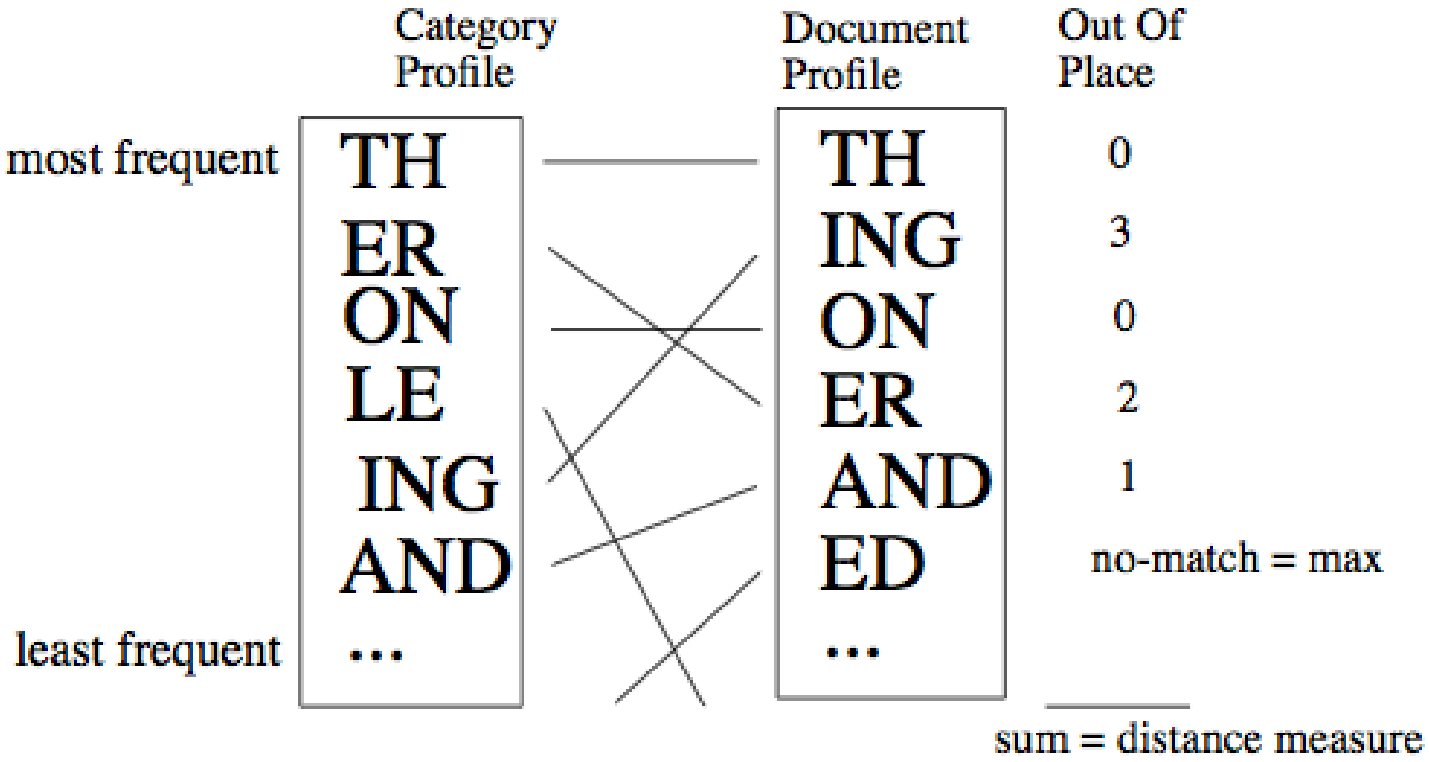


Modèles (simples) de langue

Fichiers de modèle de langue pour MGUESSER

français		**anglais**		**allemand**	
seq	freq	mot	freq	mot	freq
_	4,762,268	_	8,097,193	_	7,119,158
e	3,227,901	e	4,757,841	e	6,188,609
s	1,736,708	t	3,450,856	n	3,781,083
a	1,722,683	o	3,181,965	i	2,867,838
t	1,573,003	a	2,910,346	r	2,540,532
i	1,544,233	n	2,617,886	s	2,085,127
n	1,451,396	i	2,601,399	t	2,047,798
r	1,395,479	s	2,330,971	h	1,939,960
u	1,343,622	r	2,232,821	a	1,932,605
o	1,262,006	h	2,157,803	d	1,796,659
l	1,167,742	l	1,423,346	en	1,488,315
e_	1,105,484	d	1,405,996	u	1,388,799
d	732,432	e_	1,340,805	l	1,319,841
s_	709,985	_t	1,120,482	n_	1,299,079
t_	662,637	th	1,051,445	er	1,266,324
m	591,466	u	988,874	c	1,241,121

Comparer les distributions



$$d(a,b) = \Sigma_s |r_a(s) - r_b(s)|$$

Mise en oeuvre

Il était grilheure; les slictueux toves Gyraient sur l'alloinde et vriblaient

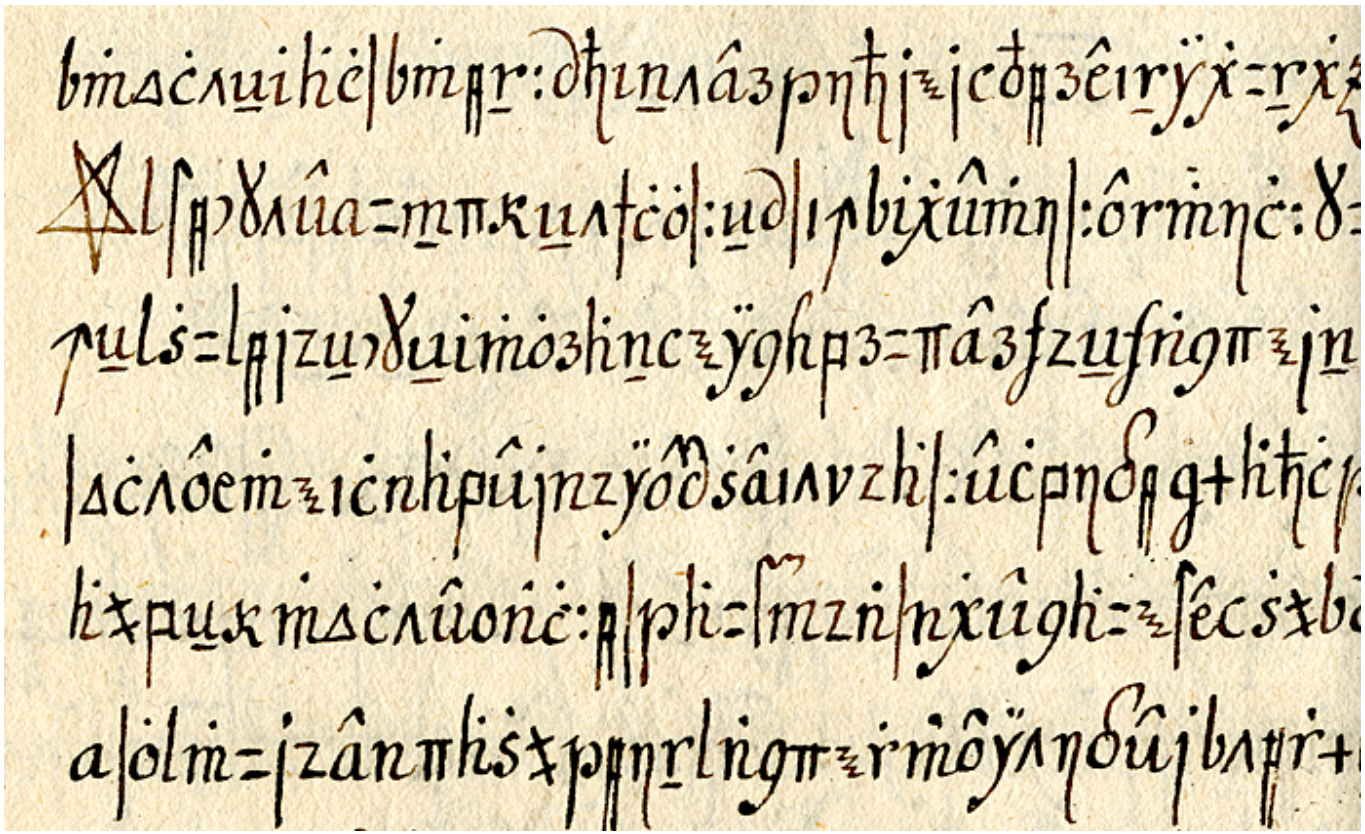
seq	freq
_	10
e	9
i	8
l	8
t	7
r	5
a	4
u	4
s	4
ai	3
n	3
t_	3
ient	2
ent	2
ien	2
ri	2

```
paste fr.latin1.mdl msg.mdl | perl ./ngram_diff.pl
```

langue	distance
fr	26,832
br	29,262
af	29,506
ca	29,576
es	29,624
no	29,656
ca	29,874
nl	30,030
la	30,036
da	30,152
ro	30,452
de	30,458
is	30,530
af	30,560
it	30,648
en	30,694

Application: le code Copiale

En 2011, Kevin Knight et ses collègues déchiffrent le code **Copiale**, utilisé dans un manuscript de 105 pages (∼ 75Kchar), datant de 1760-1780
`http://stp.lingfil.uu.se/~bea/copiale/`



Stratégie

Mise au point d'une convention d'écriture (translitération) et mise sous forme électronique.

```
## PAGE 3
" bar zzz inf gs i bar tri c ns arr v z uh g uu q ... longs
" bar o. b bar tri c. : n z o. arr l pi ns oh bar zzz bar n
" b s. ni r gam grr ns ru ih plus gs g h. bas hd lam hd zzz
" p. ki bar : oh three x z uh plus car eh nu y.. del uu a ba
" arr uu b lip uu h mal grr r. d sqi c. iot bas hd lam ki mu
" k s. ni plus p grl sqp zzz j sqi c. ni grc o sqp ih inf gs
" m ... x. ih lam hd h. g u z h. grc c ns arr b pi eh three
" ds bar o. l lam j zzz ni a ki mu pi f m. zzz ru : b uh bar
```

Comparaison avec les distributions pour diverses langues:

- pas un code par substitution
- légère proximité avec l'allemand (cohérent avec d'autres indices)

Chiffre homophonique

Les chercheurs font l'hypothèse d'un chiffre homophonique,
cherchant à masquer les probabilités des caractères:

- un caractère c à forte fréquence f peut être encodé par un caractère x dans un ensemble $\{x_1, \ldots, x_n\}$, n proportionnel à f
- utilisé pour les messages type D (calcul entropie)

Ce type de codage:

- masque les distributions sur les caractères
- mais est imparfait sur les séquences de caractères,
  en particulier sur les séquences impliquant une lettre rare
  exemple: qu en français

Succès

Code Copiale = chiffre homophonique sur de l'allemand manuscript d'initiation d'une société secrète

Plain	Cipher
A	
Ä	
B	
C	
D	
E	
F	
G	
H	
I	
J	
K	

Plain	Cipher
L	
M	
N	
O	
Ö	
P	
R	
S	
T	
U	
Ü	
V	

Plain	Cipher
W	
X	
Y	
Z	
SCH	
SS	
ST	
CH	
repeat	
EN / EM	
space	

Plain	Cipher
Logograms	

Plan

Les données

Récupération des oeuvres sur Gutemberg (en format UTF8)
`http://www.gutenberg.org`

- Stendhal
  - Le rouge et le noir (1830, 212Kmots)
  - La chartreuse de Parme (1839,219Kmots)
- Jules Vernes
  - Voyage au centre de la terre (1864, 87Kmots)
  - 20000 lieues sous les mers (1870, 175Kmots)
  - Le tour du monde en 80 jours (1873, 100Kmots)
- Gaston Leroux
  - Le mystère de la chambre jaune (1907, 109Kmots)
  - Le fauteil hanté (1909, 66Kmots)
- Maurice Leblanc
  - Arsène Lupin gentleman-cambrioleur (1907, 73Kmots)
- Marcel Proust
  - Du côté de chez Swann (1913, 201Kmots)
  - Le côté de Guermantes (1921-22, 85Kmots)

Extraction du vocabulaire

Découpage (naïf) en token: blancs, ponctuations, apostrophes (devant voyelles)

> perl ./analyze.pl pg13765.l1.txt

Du côté de . . .

mot	#occ	freq (%)
,	13,693	6.80
de	7,734	3.84
.	4,485	2.23
la	3,846	1.91
à	3,603	1.79
et	3,491	1.73
que	3,107	1.54
le	2,945	1.46
il	2,803	1.39
qu’	2,747	1.36
l’	2,476	1.23
un	2,462	1.22
d’	2,455	1.22
les	2,276	1.13

20000 lieues . . .

mot	#occ	freq (%)
,	13,912	7.92
.	7,860	4.48
de	6,238	3.55
le	3,243	1.85
et	3,066	1.75
la	2,958	1.68
à	2,762	1.57
les	2,336	1.33
l’	2,011	1.14
des	1,968	1.12
un	1,708	0.97
que	1,556	0.89
d’	1,493	0.85
–	1,432	0.82

Comparer les distributions

Comparer les variations de distributions de n mots les plus fréquents communs ($\sim$ *mots vides*)

, de . la à et que le il qu' l' un d' les qui une en pas ne des dans était pour n' du ce se s' est

Recherche d'une distance entre les rangs de ces mots

$$\text{rank-distance}(d_a, d_b) = \Sigma_w |r_a(w) - r_b(w)|$$

D'autres mesures envisageables: coefficient de corrélation de Spearman, coefficient de Kandall

Matrice des distances

Matrice des distances (rank-distance) pour $n = 50$

> perl ./spearman.pl *.voc

	Du Côté de Chez ...	La Chartreuse ...	Le mystère de ...	Le fauteuil hanté	Arsène Lupin ...	Tour Du Mond 80 ...	Voyage au Centre ...	20000 Lieues ...	Le Rouge et le ...	Le Côté de Guerma
Du Côté de Chez ...	0	62	106	92	84	108	120	118	68	32
La Chartreuse ...		0	100	92	84	78	100	90	36	66
Le mystère de ...			0	68	100	122	136	122	100	112
Le fauteuil hanté				0	76	108	134	122	88	100
Arsène Lupin ...					0	84	88	88	84	82
Tour Du Mond 80 ...						0	72	62	86	112
Voyage au Centre ...							0	46	104	102
20000 Lieues ...								0	98	102
Le Rouge et le ...									0	72
Le Côté de Guermantes										0

Regroupement en *cluster*

Regrouper les oeuvres proches en distance pour former des clusters

Utilisation d'un algo de Regroupement Hierarchique Agglomératif

1. chaque oeuvre forme un cluster
2. à chaque étape, on regroupe les 2 clusters les plus proches

$$(c_1^\star, c_2^\star) = \underset{c_1, c_2}{\operatorname{argmin}} \frac{\sum_{a \in c_1} \sum_{b \in c_2} d(a,b)}{|c_1| . |c_2|}$$

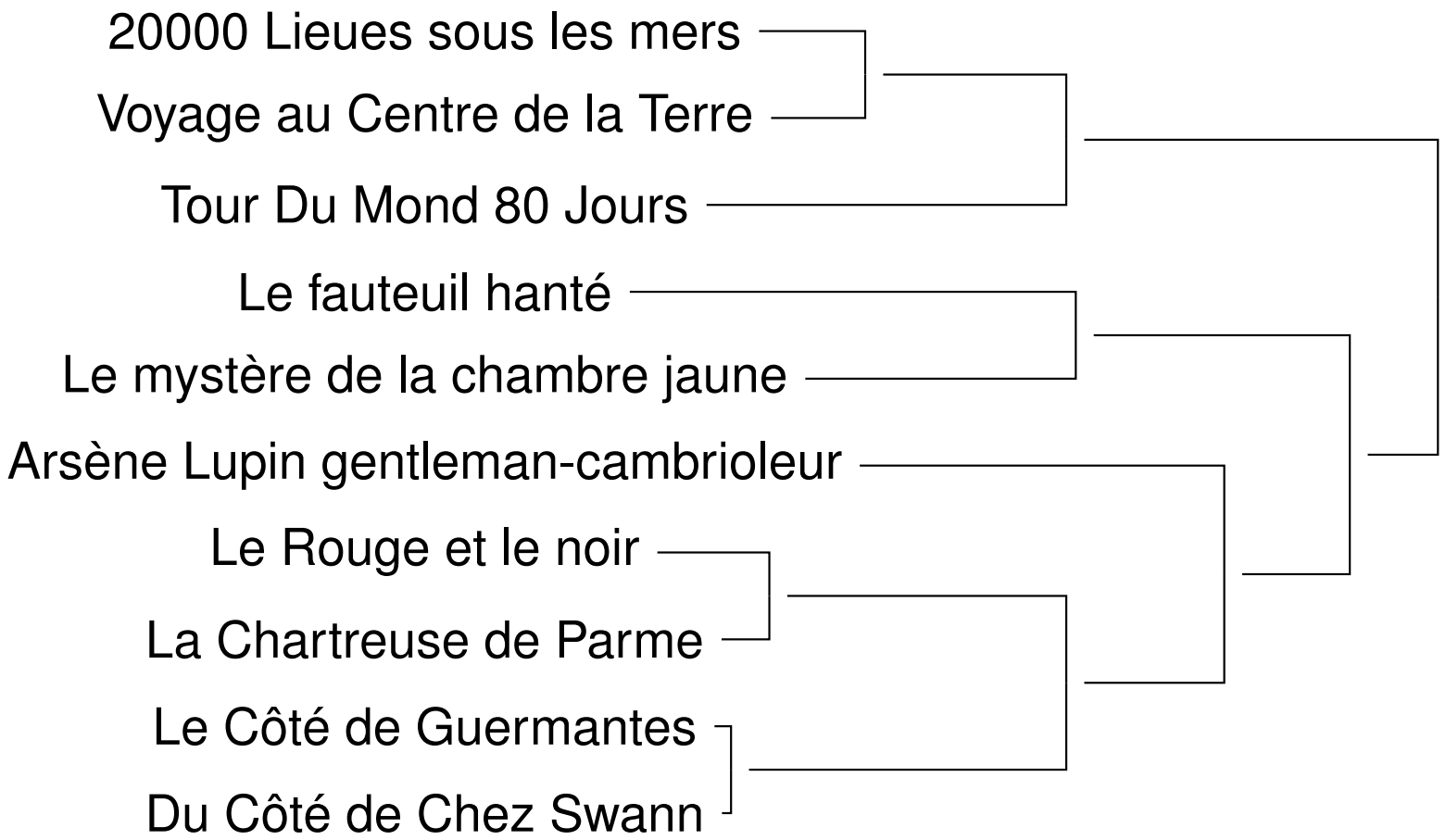
3. on stoppe quand il ne reste plus qu'un cluster

De nombreux autres algorithmes de regroupement possibles

Regroupement hiérarchique $\Longrightarrow$ arbre
visualisation sous forme de dendogramme

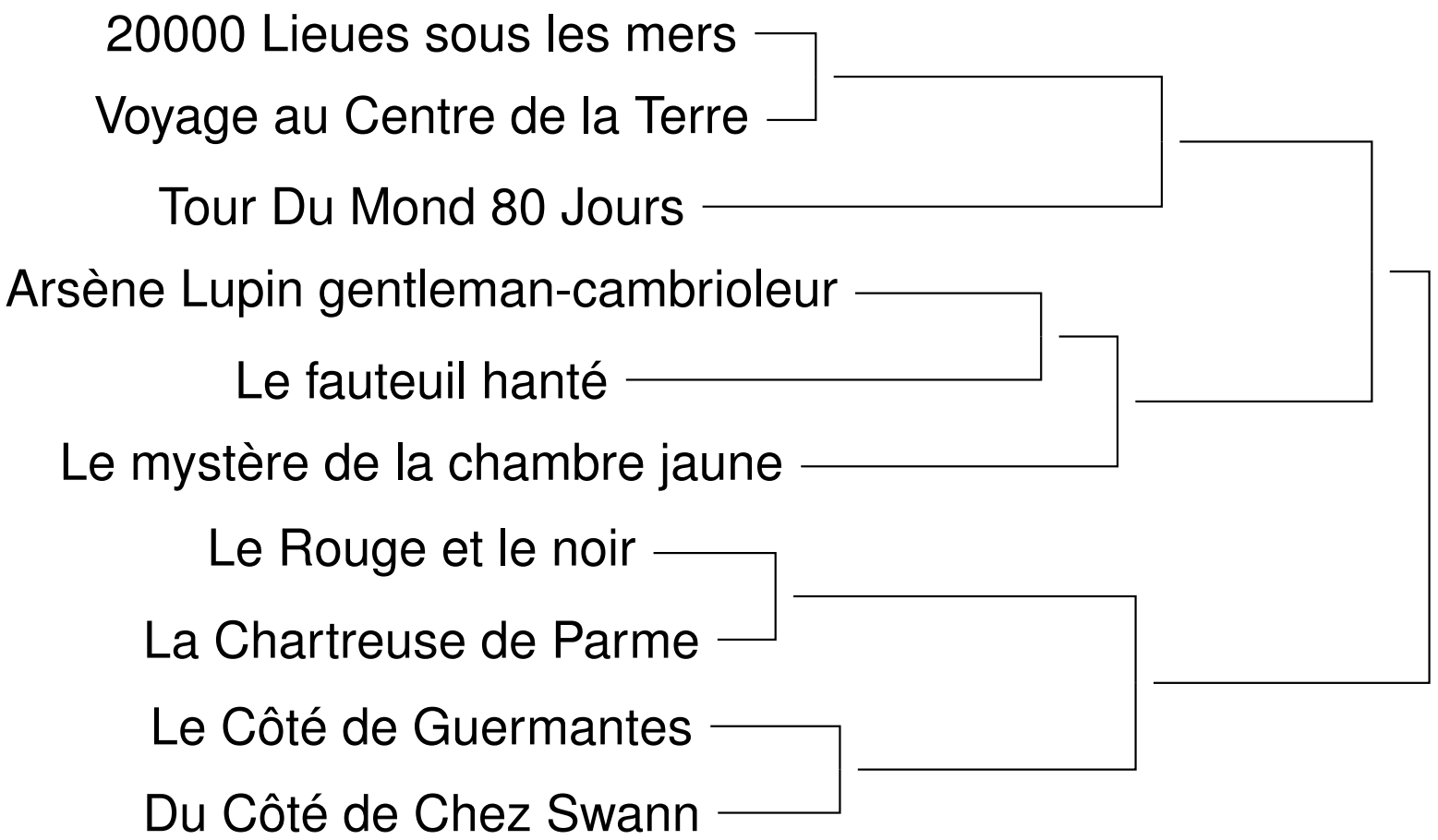
Regroupement (50)

, de . la à et que le il qu' l' un d' les qui une en pas ne des dans était pour n' du ce se s' est



Regroupement (80)

, de . la à et que le il qu' l' un d' les qui une en pas ne des avait lui dans était pour je n' comme plus du ce se s' si est son par avec au sa sur mais ! cette a



Références

- *Rank Distance as a Stylistic Similarity*
  Marius Popescu & Liviu P. Dinu
  point de départ pour cette expérience

- *Inter-textual distance and authorship attribution corneille and moliere*
  Labbé, Cyril and Dominique Labbé. 2001.
  Journal of Quantitative Linguistics, 8(3):213-231.

Linguistique Légale

Forensic Linguistic : International Association of Forensic Linguists (`http://www.iafl.org/`)

Objectifs (en autres)

- Identifier l'auteur d'un message
- Caractériser les cas de plagiat

Pour aller plus loin

Existence de compétitions sur ces tâches (PAN CLEF 2013)
`http://www.uni-weimar.de/medien/webis/research/events/pan-13/pan13-web/about.html`

Utilisation d'un ensemble plus large de traits

- fréquence lettres et séquences
- fréquence ponctuations
- fréquence préfixes et suffixes de mots
- fréquence mots et séquence de mots
- **fréquence mots vides** (et séquences)
- hapax (mots uniques)
- longueur des mots, phrases, paragraphes
- fréquence catégories syntaxiques (noms, verbes, . . .) et séquences
- traits syntaxiques, discursifs, . . .

Plan

Utilisation de modèles de langue

Déjà exploré lors du calcul d'entropie:

- n-grammes de caractères

$$p(c_{n+1}|c_1 \cdots c_n)$$

- n-grammes de mots

$$p(w_{n+1}|w_1 \cdots w_n)$$

Calculés sur de très gros corpus,
avec lissage des modèles (réseaux de neurones)

Étant donné un modèle et un texte, propose les continuations les plus probables
auto-adaptation du modèle à l'auteur (SWIFTKEY sur les smartphones)

Démos en ligne:

- `https://www.cs.toronto.edu/~ilya/fourth.cgi`

  The main problem is that a spectacular metal rocket can be ignored mere sights and conventional internal fields. In many p

Expérience sur le français

Exploitation des données consultées pour le calcul d'entropie

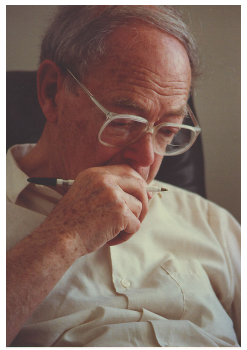
```
shell> cat pg13765.l1.txt | perl ./entropy.pl 8 4
...

> 100 il se précipite vers
il se précipite vers le pavillon m'empêcher son poste
d'observation de la hauteur. Qui dit: «Joseph Rouletabille qui
con

> word 20 il pense que
il pense que c'est le «diable» ou la «Bête du Bon Dieu», la mère
Agenoux, une vieille sorcière de Sainte-Geneviève- des-Bois, son
miaulement
```

Plan

Le sens émerge des usages

Meanings of words are (largely) determined by their distributional patterns (Harris 1968)

You shall know a word by the company it keeps (Firth 1957)



En pratique, l'ensemble des contextes pour un mot résumé par un vecteur
principe: des mots proches sémantiquement ont des vecteurs proches

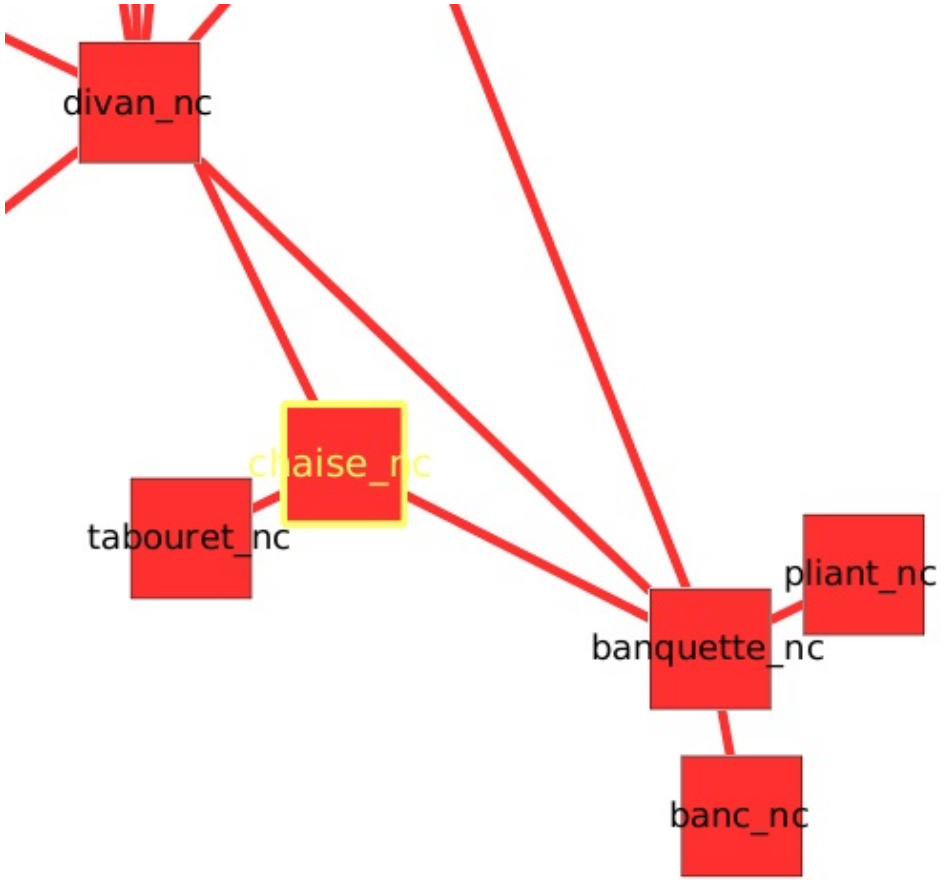
Les vecteurs sont affinés par apprentissage
(réduction des dimensions, réseaux de neurones)

Compter et collecter des dépendances syntaxiques

<governor>	<rel>	<governee>	<freq>
chaise_nc	et	table_nc	235
asseoir_v	sur	chaise_nc	227
chaise_nc	modifieur	long_adj	168
chaise_nc	de=	poste_nc	115
tomber_v	sur	chaise_nc	103
chaise_nc	modifieur	musical_adj	102
se_asseoir_v	sur	chaise_nc	93
prendre_v	cod	chaise_nc	87
chaise_nc	modifieur	électrique_adj	82
chaise_nc	modifieur	vide_adj	80
chaise_nc	à=	porteur_nc	80
dossier_nc	de	chaise_nc	78
avoir_v	cod	chaise_nc	71
table_nc	et	chaise_nc	62
chaise_nc	de=	paille_nc	56

Réseau de mots

Rapprochement des mots en fonction de la similarité de leurs contextes (*Markov Clustering*)



Divers réseaux de mots accessibles en ligne sous LIBELLEX:
`http://alpage.inria.fr/Lbx`
`login`: guest `passwd`: guest

Raisonnement par analogie

Certaines opérations sur le langage capturables par des opérations élémentaires sur les vecteurs.

Raisonnement par analogie:
logiciel VEC2WORD sur des vecteurs construits sur Wikipédia
`https://code.google.com/p/word2vec/`

Requête: **-homme +femme +fils** =

Raisonnement par analogie

Certaines opérations sur le langage capturables par des opérations élémentaires sur les vecteurs.

Raisonnement par analogie:
logiciel VEC2WORD sur des vecteurs construits sur Wikipédia
`https://code.google.com/p/word2vec/`

Requête: **-homme +femme +fils** =
fille père soeur veuve . . .

Raisonnement par analogie

Certaines opérations sur le langage capturables par des opérations élémentaires sur les vecteurs.

Raisonnement par analogie:
logiciel VEC2WORD sur des vecteurs construits sur Wikipédia
`https://code.google.com/p/word2vec/`

Requête: **-homme +femme +fils** =
fille père soeur veuve . . .

Requête: **-homme +femme +roi** =
princesse **reine** patriarche prince . . .

Reqête: **-roi +reine +prince** =
princesse soeur mère duchesse . . .

Démo en ligne sur `http://www.thisplusthat.me/`

Conclusion

- Des méthodes très simples permettent déjà de découvrir des caractéristiques intéressantes du langage
  d'autant plus que des données textuelles sont facilement accessibles
- on observe bien la richesse des langues, mais aussi une plus grande prédictibilité que soupçonnée
- base des approches statistiques actuelles par des méthodes d'apprentissage non ou faiblement supervisé
- liens entre
  - linguistique
  - théorie de l'information
  - cryptographie
  - compression

Merci