Ajuster corpus et objectifs

Serge Fleury & Benoît Habert

http ://www.cavi.univ-paris3.fr/ilpga/ilpga/sfleury/ - fleurys [@] noos [.] fr

http ://www.limsi.fr/Individu/habert - habert [@] limsi.fr

Plan

Objectif : décrire un mot/une notion **sur** corpus

Plan

Objectif : décrire un mot/une notion **sur** corpus

Face à la disette dans la profusion

- déluge de données électroniques
- pauvretés de la richesse
- prédire/picorer/modéliser
- intuition et attestation
- sonder la langue

Plan

Objectif : décrire un mot/une notion **sur** corpus

Face à la disette dans la profusion

- déluge de données électroniques
- pauvretés de la richesse
- prédire/picorer/modéliser
- intuition et attestation
- sonder la langue

laïcité et 'laïcité' : démarche

- un opportunisme relatif
- une sémantique distributionnelle et « mesurée »

Visages de la laïcité 1/2

Le Monde du mercredi 27 octobre 2004
Pour la première fois dans l'histoire des forces armées britanniques, la Royal Navy vient d'autoriser un marin à pratiquer sa religion, le satanisme. Le ministère de la défense a justifié sa décision par son souci, en tant qu'employeur, de respecter l'« égalité des chances » et de ne pratiquer aucune discrimination fondée sur les croyances de ses recrues. [. . .] En l'autorisant à pratiquer son culte en mer, son capitaine a considéré . . . que cela ne porterait atteinte « ni à l'efficacité opérationnelle du navire ni au bien-être général de l'équipage ». Il a quand même été demandé au marin de ne « rien faire qui puisse choquer » et de ne pas, par exemple, « se livrer à certains rites comme boire le sang d'animaux fraîchement décapités ».

Laïcité et données électroniques

- La laïcité comme notion joue un rôle important (conflictuel) dans l'histoire française depuis au moins 2 siècles
  - séparation église (catholique) / état (loi de 1905)
  - séparation politique / religieux (échec du MRP après-guerre)
- La laïcité est redevenue plus présente récemment (le « voile »)
- Angles d'attaque
  - lexicographique (*laïcité*) /notionnel ('laïcité')
  - en synchronie/en diachronie
  - appuyée ou non sur des données électroniques

Devant moi, le déluge. . .

- Dix ans déjà. . .

. . . *la lexicographie entre désormais dans une ère où les corpus de plus de 100 millions de mots seront monnaie courante*
[Church *et al.*, 1994]

- Ce constat vaut d'abord pour l'anglais
  - [Pantel et al. 2004] 1.5 **milliards** de mots pour découvrir des relations d'hyperonymie

Ressources disponibles 1/2

- Corpus monolingues
  - « bruts » (simples mots)
  - annotés
    - étiquetés et lemmatisés ;
    - arborés ;
    - marqués pour : sens lexicaux, dialogue, co-référence. . .
- Corpus multilingues
  - alignés (en rapport de traduction) pour phrases, groupes de mots ;
  - comparables (textes similaires).
- Lexiques monolingues ou multilingues

Ressources disponibles 2/2

- Outils d'annotation manuelle ou automatique
  - aide à la transcription d'oral ;
  - étiquetage morpho-syntaxique
  - dépendances syntaxiques
- Outils d'exploration
  - concordances ;
  - classement thématique ;
  - proximité entre textes ;
  - . . .

Richesses. . .

- L'ère des *méga-corpus* [Kennedy 98], « réservoirs » à corpus
  - 1979 corpus Brown 1 million de mots
  - 1995 BNC (*British National Corpus*) 100 millions de mots
- Ordres de grandeur
  - 5 années du *Monde* $\approx$ 100 millions de mots $\approx$ 1 000 romans $\approx$ 10 000 heures de conversation
  - $\approx$ 3,8 milliards de mots en français sur le Web en mars 2001 [Kilgarriff et Grefenstette, 2003]

Pauvretés de la richesse

- Dans le BNC, l'essentiel des mots apparaissent moins de 50 fois [Kilgarriff et Grefenstette, 2003]
- [Gasiglia 04] *Le Monde* 1997 et 1998 fournit 3 586 articles sur le footbal dont seulement 20% sont pertinents pour les emplois visés de *passer*
- 14 millions de mots du *Monde* (numéros entiers extraits aléatoirement des années 1987, 1989, 1991, 1993, 1995) : pas d'emploi de 'trahir' dans les 1 345 occurrences du verbe *vendre*

Le syndrome de l'ivrogne au réverbère

- Risque de privilégier les corpus « opportunistes », ie disponibles mais ne s'accordant pas forcément avec les objectifs de la recherche.
  - Ex. versions numériques de grands journaux et variation selon les sections et les genres [Illouz *et al.* 1999]
  - Ex. Frantext, langue littéraire et périodes représentées
- « Discrimination positive » – sur-représenter les emplois visés
  - [Vaguer 2004] *dans* « de coïncidence » (*il s'est trompé dans l'administration du médicament* $\equiv$ 'du fait qu'il a administré le médicament)
  - [Gasiglia 2004] « corpus à haut rendement » de transcriptions de commentaires de matchs
- Opposition (Biber) linguistique basée sur des corpus / « tirée » par des corpus

Tirer des bords

- Les corpus disponibles, leurs annotations, les outils accessibles ne répondent jamais exactement aux objectifs de telle recherche
  - Ex. [Leroy 2004] s'appuie sur des suites plates d'étiquettes morpho-syntaxiques et de lemmes pour repérer de possibles antonomases et non sur des syntagmes : identifier la frontière droite est difficile.
- Composer avec le caractère nécessairement imparfait des ressources
  - Ex. [Leroy 2004] L'étiqueteur utilisé « oublie » des noms propres et baptise à tort noms propres certains mots.
- Le TAL et les industries de la langue ne fourniront jamais des données annotées « pures » et sans erreur

Disponibilité des ressources ?

- Du « libre » au protégé
  - ressources libres : WordNet, WinBrill ;
  - commercialisation : Cordial `http://www.synapse.com` ;
  - prototypes de recherche ;
  - produits industriels.
- Accessibilité effective
  - gamme de prix ;
  - consultation/utilisation : ex. consultation du *Trésor de la langue française* entrée par entrée (`http://www.atilf.fr`) / connaître le nombre d'acceptions de chaque entrée nominale ;
  - complexité plus ou moins grande de prise en main.

Souvent langue varie. . .

Les linguistes de corpus ont eu tendance à travailler avec de larges volumes venant de pseudos-genres flous du type ‘textes journalistiques’, mélangeant les écrits de nombreuses personnes et provenant d’endroits variés.

[Manning 2003]

6 sections du *Monde Parole*

7 millions de mots (extrait de 14 millions de mots de numéros du *Monde* tirés aléatoirement des années 87, 89, 91, 93 et 95 [Illouz et al. 1999]

Rubrique	sous-emplois	sur-emplois
ARTs	dét, prép, V	N, adv, pron, conj
ECO(nomie)	V, pron, adv, conj	dét, prép
EMS (Education, Médecine, Société)	nom, adv	conj
ETR(anger)	pron, adv, conj	adj
INF(o. géné.)	adj	V, pron, adv
POL(tique)	nom, dét, prép, adj	V, pron, adv, conj

Le Monde Parole : traits linguistiques

Trait	ART	ECO	EMS	ETR	ING	POL
NonPers.	**+08**	-11		-13	**+04**	**+08**
Embrayeur	**+16**	-16	-03	-17	**+07**	**+08**
CCoord	**+07**	-02	**+11**	-04		-03
CSub	**+02**	-02		-04		**+06**
Part. nég.	-03	-09	-02	-05	**+05**	**+24**
Adv degré	**+06**	**+03**		-03		-04

Genres et catégories 1/2

LOB [Biber 1993]

- Mots « pleins »

Forme	cat.	fiction %	« exposition » %
trust	N	18	85
	V	82	15
rule	N	31	91
	V	69	9
major	titre	69	11
	A	31	85

Genres et catégories 2/2

- Mots « outils »

Forme	cat.	fiction %	« exposition » %
major	titre	69	11
	A	31	85
that	dét.	37	17
	conj.	45	69
	rel.	14	11
before	prép.	30	54
	sub.	48	32
	adv.	22	14

Hétérogénéité d'un genre

[Biber & Finegan 1994]

- Articles scientifiques (New England Journal of Medicine, Scottish Medical Journal)
- Contrastes entre les parties canoniques
  - Introduction et conclusion : 1ère pers., complétives en *that*
  - Méthodes : passif sans agent, priorité au passé sur le présent
  - Discussion : modaux de possibilité et présent

Une pauvreté destinée à perdurer

. . . je ne crois pas qu'il puisse y avoir un corpus, si grand soit-il, qui puisse fournir de l'information, pour l'anglais, sur tous les secteurs du lexique et de la grammaire que je veux explorer . . .

[Fillmore 1992, p. 35]

Inévitables compromis

Il n'y a pas de réponse facile au problème d'obtenir des données suffisantes exactement du bon type : la langue change selon le temps, l'espace, la classe sociale, les méthodes d'obtention [methods of elicitation], *etc. Il n'est pas possible d'obtenir une grande quantité de données (ou du moins une collection où le phénomène visé est fortement présent) sans se préparer à piétiner allègrement au moins une de ces dimensions de variations.*

[Manning 2003]

Prédire / picorer / modéliser

[Gross 1988 ; 1996] Restrictions sur les transformations dont une séquence serait théoriquement passible comme mesure de son degré de figement et comme pierre de touche des 'expressions figées' ou 'mots composés'
Ex. de *guerre froide*
la paix froide (J. Ellenstein), *la Guerre-qui-est-restée-froide* (J. Roubaud), *guerre à froid* (Google)

Prédire

Propriété		*exemple*	Google
prédicativité	-	*la guerre est froide	5
adj. nominalisé	-	*la froideur de la guerre	4
commutation adj.	-	*la guerre chaude	1
sing./plur.	+	les guerres froides	383
adv. adj.	-	*la guerre très foide	39
adj. coordonné	-	*guerres froide et chaude	1
reprise par nom	+	cette guerre. . .	?
commutation nom	-	*la paix froide	283
adj. $\rightarrow$ *de* nom	-	*la guerre de froideur	0

Sur la Toile

1. *Cette **guerre** est **froide** car il n'y a pas d'affrontement direct entre les deux grands*. . .
2. *La **froideur** de la **guerre** est peut-être en cause*.
3. *Aux postes les plus sensibles de la CIA. . ., il est l'un des protagonistes de la **guerre** très **froide** qui, de 1985 à la chute du rideau*. . .
4. ***guerre** très **froide*** : nombreuses citations dans des explications de jeux vidéo
5. *Cette radicalisation de l'offensive contre l'informatique libre dessine une sorte de nouvelle **guerre froide** et sans pitié*
6. ***guerre** chaude et paix **froide***
7. *année triste de **guerre,** de **froid** et de privations*

Surf sans conscience. . .

- L'accès aux documents est médiatisé par le moteur de recherche
  - un moteur indexe une partie variable de la Toile
  - la partie accessible évolue au fil du temps pour un même moteur
  - indexation $\approx$ déformation (casse, accents, ponctuation)
  - un même événement peut être plusieurs fois répertorié et donner lieu à de multiples attestations (*Les guerres froides* $\equiv$ 2 livres – Y. Reynaud ; V. Lou)
- Des OVNI (Objets Verbaux Non identifiés), orphelins de leur contexte

Mesurer la flexibilité 1/4

[Barkema 1993 ; 1994] Recherche des variations d'expressions dans le corpus de Birmingham (20 millions de mots)

Schéma	#	*exemple*
D *c w*	111	
D A *c w*	3	the melting Cold War
D *c w* prop.	2	the Cold War that existed. . .
D *c w* SP	2	the cold war between. . .
D *c w* part. passé	1	the Cold War won by Europeans
D A *c w* part. passé	1	the awkard cold war thought up b
D Adv *c w* SP	1	a not-so-cold war against Kadda
D N *c w*	1	the world Cold War
D *c* A *w*	1	a kind of cold civil war
D *c* Cc A A *w* prop.	1	a period of cold and hot civil war

Mesurer la flexibilité 2/4

Utilisation du corpus arboré de Nimègue (130 000 mots, 16 183 SN relevant de 1 736 patrons syntaxiques)

fréquence théorique des réalisations A de *x* =

$$\text{occ. de } x \;*\; \frac{\text{occ. de la réalisation A dans les exp. libres}}{\text{occ. des patrons de } x \text{ et de ses variantes dans les exp. libres}}$$

Mesurer la flexibilité 3/4

Schéma	% *attendu*	% *réel*	*écart*
dét. *cold war*	39,64	89,52	**+49,88**
dét. *cold war* prop.	6,12	1,60	-4,52
dét. *cold war* SP	15,52	1,60	**-13,92**
dét. *cold war***s**	19,87	0,00	**-19,87**
dét. *cold war***s** SP	2,30	0,00	-2,30

Mesurer la flexibilité 3/4

Limites des expériences de Barkema

- la flexibilité des syntagmes libres comme des expressions ‘toutes faites’ varie peut-être selon les registres (quotidiens / livres d’histoire pour guerre froide)
- le corpus étalon est distinct du corpus où figurent les variations
- le patron dont les variations sont examinées – adjectif nom commun – est trop grossier
- les sous-classes d’adjectifs et de noms influent sur les combinaisons de l’ensemble
- mesurer sur des sous-classes « diluerait » les constats quantitatifs et les fragiliserait

linguistique de l'attesté / du possible

Ces deux linguistes [le linguiste de corpus et le linguiste dans son fauteuil – corpus linguist *et* armchair linguist*] ne se parlent pas très souvent, mais quand ils le font, le linguiste de corpus dit au linguiste dans son fauteuil 'Qu'est-ce qui pourrait me faire penser que ce que vous me dites est vrai ?', et le linguiste dans son fauteuil répond au linguiste de corpus 'Qu'est-ce qui pourrait me faire penser que ce que vous me dites est intéressant ?'*

[Filmore 92, p. 35]

Complémentarité attesté / possible

[l'introspection et le corpus] *découpent le champ des recherches linguistiques en deux domaines, qu'il est commode de baptiser schématiquement 'linguistique de bureau' et 'linguistique de terrain', et dont aucun ne peut légitimement être présenté comme incarnant à lui seul LA linguistique : éventuellement complémentaires, voire présentant certaines intersections, ces deux linguistiques ne peuvent pas avoir globalement le même objet. Si l'introspection peut repérer certaines variations dans les pratiques langagières, elle est impuissante à décrire leur distribution dans la population : le social lui échappe par définition. Inversement, la linguistique de terrain, qui s'efforce de prendre en charge les aspects sociaux des variations langagières, privilégie par force certains secteurs de la description linguistique.*

[Corbin 1980, p. 121]

Représentativité : intuition et attestation

L'exemple n'importe pas par sa particularité, mais par sa généralité. Il faut donc admettre qu'il est le nom d'une classe. . . Comment passe-t-on de la phrase particulière aux phrases qui lui sont semblables ?

[Milner 1989, p. 52]

. . . l'intervention grammaticale majeure a lieu quand une donnée matériellement attestée est jugée linguistiquement impossible.

[Milner 1989, p. 55]

La règle est une hypothèse sur les faits, les faits contiennent aussi bien du possible que de l'impossible.

[Auroux 1998, p. 197]

Les corpus de référence 1/2

Un corpus de référence est conçu pour fournir une information en profondeur sur une langue. Il vise à être ***suffisamment grand*** *pour représenter* ***toutes les variétés pertinentes*** *de cette langue et son vocabulaire caractéristique, de manière à pouvoir servir de base à des grammaires, des dictionnaires et d'autres usuels fiables*.

[Sinclair 1996]

Les corpus de référence 2/2

- Tradition anglo-saxonne de corpus « panachés »
  - (échantillonnage)
  - prise en compte de genres ou registres diversifiés
- Exemples
  - Corpus Brown : 15 genres, échantillons de 2 000 mots ;
  - British National Corpus : oral (10% – associant interactions typiques et conversations spontanées), écrit (fiction et textes informatifs)

Sonder une population / la langue

- Sondage de convenance : sélectionner les individus les plus accessibles
- Sondage par quotas : respecter dans l'échantillon certaines proportions de la population
- Sondage aléatoire

- Les corpus « opportunistes » se rapprochent des sondages de convenance
- Peut-on pour la langue utiliser des quotas ou des sondages aléatoires ?

Profiler les textes

- L'existence de « réservoirs à corpus » implique de pouvoir
  - évaluer leur homogénéité interne
  - dégager des sous-parties homogènes
  - assembler de manière raisonnée des documents

[Simonin-Grumbach 1975][Sueur 1982]
[Illouz et al. 1999][Malrieu & Rastier 2001][Beauvisage 2001][Folch et al. 2000][Beaudouin et al. 2002]. . .

- A défaut de pouvoir disposer de corpus représentatifs, il faut savoir ce que représentent les corpus dont nous disposons
- corpus : collection de données langagières qui sont sélectionnées et organisées selon des **critères linguistiques et extralinguistiques explicites** pour servir d'**échantillons d'emplois déterminés** d'une langue

Au fond de l'inconnu, trouver du nouveau

. . . *chaque corpus que j'ai eu l'occasion d'examiner, si petit soit-il, m'a montré des faits* [taught me facts] *que je ne pourrais pas imaginer découvrir d'une autre manière.*

[Fillmore 1992, p. 35]

• Opposition (Biber) linguistique basée sur des corpus / « tirée » par des corpus

Exemple [Biber et al. 1999] les « blocs lexicaux » – cf. les segments répétés de [Salem 1987] – retours récurrents de 3 à 6 mots, distincts des expressions idiomatiques et qui jouent un rôle structurant dans la conversation en particulier (*I tell you what, know what I mean*, etc.)

Arche de Noé ?

Un lexicographe ressemble à quelqu'un se tenant sous les chutes du Niagara avec un pluviomètre comme seul instrument de mesure, tandis que les données pertinentes coulent sur lui en torrents démesurés
[Church *et al.* 1994]

Une sémantique outillée ?

L'une des particularités des sciences du langage, . . . c'est que le langage est sans médiation à disposition du lecteur : je puis produire, à volonté, des phrases, les tronquer, y introduire tel élément que je choisis, etc. Il se pourrait que ce soit le seul exemple d'une manipulation sans instrument, du moins le seul qui se soit maintenu dans un état développé d'une discipline scientifique.

[Auroux, 1998, p. 170]

[. . .] le progrès de la sémantique sera à l'avenir, au moins pour partie, lié à son automatisation ; [. . .] dans le 'Traitement automatique des langues' [. . .], il y a tout à gagner à accroître la part de la sémantique ; [. . .] les 'dictionnaires informatisés' [. . .] peuvent jouer un rôle important dans l'évolution envisagée.

[Martin, 2001, p. 11]

Laïcité : démarche

Sites : http ://netx.u-paris10.fr/habert_benoit/PluriTAL/

http ://www.cavi.univ-paris3.fr/ilpga/plurital/cours2-2004.html

- Stabiliser les intuitions de départ et mesurer les écarts
  - formulaire et synthèse correspondante
- Un opportunisme relatif
  - ne pas se faire une/la Toile
  - deux « tranches » d'un journal généraliste à 10 ans d'intervalle
  - des données de nature différente
    - « capture » d'une version en ligne
    - version électronique étiquetée/lemmatisée et dépendances syntaxiques
- Une sémantique distributionnelle et « mesurée » (mesurant)

Lai, laie – Le Petit Robert

adj. XIIe ; lat. ecclés. *laicus*, gr. *laikos*, de *laos* « peuple »
vx Laïque
Mod. *Frère lai* : frère servant ⇒ **convers**

Laïc – Le Petit Robert

→ *laïque*
1487 ; lat. ecclés. *laicus*

1. Qui ne fait pas partie du clergé, et spécialt Qui n'a pas reçu les ordres de cléricature, en parlant d'un chrétien baptisé.
2. FIG. (avec un subst. désignant normalement un religieux) « *Un saint laïque* » (Pasteur, à propos de Littré)
3. Qui est indépendant de toute confession religieuse (⇒ laïcité)

♢ Contr. Clerc, ecclésiastique. Religieux.

Laïcité – Le Petit Robert

n. f. 1871 ; de *laïc*

1. Caractère laïque
2. Principe de séparation de la société civile et de la société religieuse, l'Etat n'exerçant aucun pouvoir religieux et les Eglises aucun pouvoir politique. « *la laïcité, c'est-à-dire l'Etat neutre entre les religions* » (Renan).

Laïcité : exemple 1/2

Le Monde, janvier 1991

Chef d'un Etat qui a toujours claironné sa laïcité *, M. Saddam Hussein n'hésite pas , également , à reprendre les arguments religieux de guerre sainte qu'il a découverts avec la crise du Golfe et à parler de " bataille suprême " .*

Triplets Syntex [Bourigault & Fabre 2000]

1. <SUJ, état, claironner>
2. <OBJ, claironner, laïcité>

Laïcité : exemple 2/2

lemme	*forme*	*catégorie*
chef	Chef	NomMs
de	d’	Prep
un	un	DetMS
état	Etat	NomMS
qui	qui	ProRel
claironner	a claironné	VCONJS
toujours	toujours	Adv
son	sa	DetFS
laïcité	laïcité	NomFS

Laïcité janvier 1991 *Le Monde*

SUJ <SUJ, nouer, tendance de le laïcité>

OBJ <OBJ, adopter, laïcité> <OBJ, claironner, laïcité> <OBJ, confronter, laïcité> <OBJ, instaurer, laïcité> <OBJ, rénover, laïcité> <OBJ, restreindre, problème de le laïcité>

EPI <EPI, acquis, laïcité> <EPI, actualité, laïcité> <EPI, crime, lèse-laïcité> <EPI, désenchantement, laïcité> <EPI, histoire, laïcité> <EPI, laïcité, français> <EPI, modèle, laïcité> (2 o.) <EPI, mot, laïcité> <EPI, problème, laïcité> <EPI, question, laïcité> <EPI, symbole, laïcité> <EPI, tendance, laïcité>

sur <sur, falloir, acquis de le laïcité> <sur, insister, actualité de le laïcité>

à <à, laïcité, épreuve> <à, opposition, laïcité>

Laïcité : une famille

Le Petit Robert

laïcat (1877) ensemble des chrétiens non écclésiastiques

laïcisation (1870)

laïciser (1870)

1. rendre laïque. Pronom. « *Le sentiment religieux* [. . .] *se laïcise déjà* » (Mart. du G.)
2. organiser suivant les principes de la laïcité

laïcisme (1877) doctrine qui tend à donner aux institutions un caractère non religieux.

Onomasiologie / sémasiologie

sémasiologie partir des signes (les mots)

onomasiologie partir des « concepts », des idées

[Rey 1977] . . . il n'y a pas encore de lexicographie onomasiologique, dans la mesure où le concept est insuffisamment défini pour servir de base à une technique pratique. C'est en terminologie, et dans la mesure où le système conceptuel est stable et cohérent, que cette approche est efficace.

A. Rey [1977] défend pourtant la complémentarité des deux approches.
$\Rightarrow$ *laïcité* **et** 'laïcité'

Laïcité voisins, amis, alliés. . . 1/2

[Péchoin, 1992] Thésaurus Larousse
L'homme → la vie spirituelle → le sacré et le profane → profane

N

- Profane ; incroyance 480. – Monde, siècle. Hist. relig. : nation, païen, gentilité, paganisme 476. – Gentil, païen, laïc.
- Mondanité, sécularité ; **laïcité**
- **Sécularisation**. – Laïcisation. – Théol. : profanation, exécration.

La société → la vie collective → société et organisation politique → régime
Bipartisme, tripartisme ; multipartisme, **pluralisme**, pluripartisme. – Confessionnalisme ; laïcité. – **Théocratie**.

Laïcité : vague à l'âme…

[Reboul 1994] : « … un terme est **ambigu** si l'on peut lui attribuer plusieurs extensions au moins partiellement différentes, alors qu'un terme est **vague** si l'on a des difficultés à déterminer précisément son extension ».
Elle distingue à la suite de [Kleiber 1987] :

- vague observationnel (*Pierre est grand* – le critère à appliquer pose problème) ;
- vague subjectif (*Pierre est beau* – les critères varient selon les individus) ;
- vague multi-dimensionnel (*C'est une sorte d'oiseau* – un grand nombre de critères sont à examiner pour savoir si le terme s'applique ou non).

Laïcité voisins, amis, alliés. . . 2/2

Cooccurrences $\leftrightarrow$ dépendances (récursives)

You shall know a word by the company it keeps [Firth 1957]

[Harris 1988] *Caractériser les mots par leur sélection permet de considérer le type et le degré de recouvrement, d'inclusion et de différences entre mots par rapport à leurs ensembles de sélection.*
. . . dans la plupart des cas, la sélection d'un mot inclut un ou plusieurs domaines cohérents de sélection.