Corpus minimal et proportions d'effectifs dans une enquête de lexicométrie bilingue

1. Remarques préliminaires

La plupart des statistiques lexicales ont pour objet d'analyser des textes (surtout littéraires) et leur but est tantôt d'identifier un auteur, tantôt de caractériser une période de son activité. Ces analyses ont souvent un caractère diachronique et comparatif ce qui explique la largeur des corpus et la variété des paramètres quantitatifs (tels la fréquence, la productivité ou la répartition).[1]

Rares sont les cas où le corpus d'une statistique lexicale est pris dans un dictionnaire[2], et parmi ces dictionnaires, il n'y a pas un seul de bilingue. Cela s'explique certainement par le fait qu'une analyse traditionnelle des fréquences n'a pas trop de sens dans un dictionnaire. Plus prometteuses sont les enquêtes statistiques portant sur les rapports intérieurs de la métalangue et ses différentes catégories grammaticales et sémantiques.

La catégorie grammaticale ou sémantique est un caractère qualitatif qui, n'étant pas mesurable, peut avoir des traits assez subjectifs: les rapports obtenus ne se prêtent pas à la représentation graphique ou à la représentation dans des tables de corrélation. Ce petit exposé est un essai d'appliquer la méthode statistique à un corpus lexicographique bilingue, en représentant certains rapports catégoriels dans des tables de contingence. Le corpus de cette enquête est pris dans le manuscrit électronique du futur dictionnaire français-hongrois, en préparation au Département d'Études Françaises de l'Université József Attila (Szeged).

Le dépouillement des données a été facilité par le fait que les premiers 50 fichiers du dictionnaire ont été rédigés à l'aide d'un logiciel à système SGML (Standard Generalized Markup Language): grâce au système de codage, ces fichiers constituent une base de données.[3] Pourtant, les possibilités d'erreur sont nombreuses: c'est que la structure du système de codage ne permet pas la répétition de certains éléments à l'intérieur d'un bloc donné. Il arrive par

---

[1] Par exemple: Ch. Bernel: «La richesse lexicale de la tragédie classique: Corneille et Racine.», in *Le français moderne*, janvier 1978, 1, pp. 44-53. – É. Brunet: *Le vocabulaire de Jean Giraudoux. Structure et évolution.* Slatkine, Genève, 1981. – D. Labbé: *Le vocabulaire de François Mitterrand.* Presses de la Fondation Nationale des Sciences Politiques, Paris, 1990.

[2] Comme par exemple: É. Brunet: «L'analyse statistique du Trésor de la Langue Française.», in *Le Français Moderne*, janvier 1978, 1, pp. 54-66. – R. Martin: «Syntaxe de la définition lexicographique: étude quantitative des définissants dans le Dictionnaire fondamental de la langue française.», in *Statistique et linguistique. Actes et Colloques*, N° 15, Klincksieck, Paris, 1974, pp. 61-71.

[3] J. Pajzs: *Számítógép és lexikográfia. Az MTA Nyelvtudományi Intézete, Linguistica Series A: Studia et dissertationes 4*, Budapest, 1990. – J. Pajzs: «Készülõ szótárak mint adatbázisok», in *Nyelvtudományi Közlemények* 93 (1992-1993), pp. 161-171.

exemple assez souvent qu'à l'intérieur d'un seul bloc sémantique il y a plusieurs registres de langue ou plusieurs domaines d'emploi: tous ces éléments sous le même code.[4]

Notre enquête a donc un double objectif: nous sommes d'une part curieux de voir comment les dimensions du corpus peuvent modifier les rapports entre les différentes catégories verbales, et, d'autre part, nous voudrions savoir si l'on peut définir des rapports entre les différents ensembles de catégories verbales et les indications sémantiques des verbes qu'ils contiennent, rapports présentables dans des tables de contingence. C'est le verbe qui fournit le plus grand nombre d'informations de ce genre.

2. Problèmes d'échantillonnage: ensembles de catégories et proportions

Il est évident que le corpus doit être représentatif quant au nombre de ses éléments, et suffisant quant à ses dimensions. La condition la plus importante de la représentativité, c'est la présence de toutes les catégories examinées dans le corpus. Si l'enquête se contente de la totalité des indications sémantiques, un corpus relativement petit s'avérera suffisant. Par contre, lorsque toutes les intersections possibles des catégories grammaticales doivent être prises en considération, la nécessité d'un corpus plus vaste s'impose.

Quel est donc le corpus minimal qui contient toutes les intersections possibles des catégories verbales? D'après un dépouillement préalable, c'est un corpus d'à peu près 10 fichiers (feuilles d'auteur), pourvu qu'un fichier contienne 2000 données: 200 entrées, 800 unités lexicographiques (c'est-à-dire équivalents et exemples/traductions) et 1000 données d'ordre très différent: indications, renvois, etc.

Pour illustrer ce que nous venons de dire, voilà d'abord les catégories verbales et leurs intersections d'un seul fichier, DEV-DIM (D8):

	Occurrences
Catégories verbales	
v tr	24
v intr	3
v pron	2
Intersections de catégories	
v tr & v intr	5
v tr & v pron	8

On ne retrouve donc pas, dans ce seul fichier, les catégories et intersections de catégories suivantes qui sont pourtant plus ou moins fréquentes dans le manuscrit du dictionnaire:

v tr indir	v tr & v intr & v pron
v tr & v tr indir	v intr & v tr & v pron
v tr indir & v intr	v intr & v tr indir & v pron
v tr indir & v pron	v tr & v intr

[4] M. Pálfy: «Où en sont les travaux du nouveau dictionnaire français-hongrois?», in *Acta Romanica 15: Studia Lexicographica 3* (1995), pp. 5-12.

v intr & v tr indir — v tr & v intr & v tr indir
v intr & v pron — v tr & v intr & v pron
v tr & v tr indir & v pron — v tr & v intr & v tr indir & v pron

Explication des signes:

v tr verbe transitif
v tr indir verbe transitif indirect
v intr verbe intransitif
v pron verbe pronominal
& intersection: verbes à plusieurs catégories

Par contre, un corpus à 10 fichiers est suffisamment grand pour une gamme plus large et plus nuancée des intersections possibles (ce n'est que les combinaisons vraiment rares qui manquent):

Catégories verbales des fichiers A1 – A8, selon les occurrences

1. v tr & v pron	109
2. v tr	88
3. v pron	13
4. v intr	11
5. v tr indir	8
6. v tr & v pron & v intr	6
7. v tr & v intr	5
8. v tr & v pron & v tr indir	3
9. v intr & v tr indir	2
10. v tr & v tr indir	1
11. v tr & v intr &v tr indir	1

3. Table de contingence des effectifs dépouillés, suivant les occurrences des catégories verbales

Fichiers A1 – A8: occurrences	IDS	RCT	RDL	DDS	Total
v tr & v pron/109	348	151	130	45	674
v tr/88	193	35	53	23	304
v pron/13	35	11	5	4	55
v intr/11	17	6	6	1	30
v tr indir/8	21	9	4	1	35
v tr & v intr & v pron/6	22	6	23	0	51
v tr & v intr/5	8	0	0	5	13
v tr & v tr indir & v pron/3	3	14	3	0	20
v tr indir & v intr/2	0	2	2	0	4
v tr & v tr indir/1	2	1	3	1	7
v tr & v tr indir & v intr/1	0	1	2	0	3
	649	236	231	80	1196

Abréviations: IDS indication de sens RDL registre de langue
RCT rection DDS domaine de spécialité

Ces quatre catégories sont les plus fréquentes dans la délimitation des blocs sémantiques. L'emploi des indications de sens et des registres de langue est tout à fait naturel. Quant aux RCT, il est évident que les différentes structures sont, dans la plupart des cas, liées à des champs sémantiques diverses. Le nombre relativement élevé des DDS s'explique par le fait qu'ils se prêtent excellemment à distinguer les blocs sémantiques. Indépendamment de cela, la principale fonction de ce code est de permettre à l'utilisateur d'effectuer des recherches aussi complètes que possible dans la version électronique du dictionnaire. La simple grandeur numérique nous permet donc de *séparer très nettement les catégories v tr et v tr & v pron du reste.*

L'examen des pourcentages donne un résultat plus nuancé.

Proportions: horizontalement, c'est-à-dire en fonction du pourcentage des IDS, etc. à l'intérieur des ensembles de catégories verbales:

Fichiers A1 – A8: occurrences	IDS	RCT	RDL	DDS	Total
v tr & v pron/109	51	22	19	8	100
v tr/88	63	11	17	9	100
v pron/13	63	20	9	8	100
v intr/11	56	20	20	4	100
v tr indir/8	60	25	12	3	100
v tr & v intr & v pron/6	43	12	45	0	100
v tr & v intr/5	61	**0**	**0**	**39**	100
v tr & v tr indir & v pron/3	15	70	15	0	100
v tr indir & v intr/2	0	50	50	0	100
v tr & v tr indir/1	28	14	44	14	100
v tr & v tr indir & v intr/1	0	33	67	0	100

Dans cette table, les combinaisons fréquentes (les 7 lignes du haut) se distinguent de celles qui sont plus rares: dans le cas de ces dernières, la valeur des IDS est plus petite de celle des combinaisons plus fréquentes; par contre, la valeur des RCT et des RDL est supérieure par rapport à celle des 7 premières lignes. Les DDS sont discutables, justement à cause de la valeur très élevée des *v tr & v intr.*

Proportions: verticalement, en fonction du pourcentage des catégories verbales à l'intérieur des indications sémantiques:

Fichiers A1 – A8: occurrences	IDS	RCT	RDL	DDS
v tr & v pron/109	54	64	57	57
v tr/88	30	15	23	29
v pron/13	5	5	2	5
v intr/11	3	3	3	1
v tr indir/8	3	4	2	1
v tr & v intr & v pron/6	4	3	10	0

v tr & v intr/5	1	**0**	**0**	**6**
v tr & v tr indir & v pron/3	0	6	2	0
v tr indir & v intr/2	0	0	0	0
v tr & v tr indir/1	0	0	1	1
v tr & v tr indir & v intr/1	0	0	0	0
	100	100	100	100

Cette table paraît justifier les précédentes: ici on peut observer du même coup d'œil les deux tendances divergentes des tables présentées ci-dessus.

Remarque*:* dans ces tables de contingence, une ligne horizontale sépare les valeurs nettement divergentes; les chiffres gras indiquent les valeurs dont l'appartenance est discutable: dans le cas des RCT et des RDL c'est égal à zéro, dans le cas des DDS on a une valeur très élevée, avec, en haut et en bas, des valeurs zéro.

4. Conclusions

a) Un corpus minimal de 10 fichiers (feuilles d'auteur) peut contenir presque toutes les intersections possibles des catégories verbales, pourvu qu'un fichier contienne 2000 données: 200 entrées, 800 unités lexicographiques (équivalents et exemples/traductions) et 1000 données d'ordre très différent: indications, renvois, etc.

b) Grâce à une très simple enquête statistique on constate donc que:
   - l'importance des IDS l'emporte sur celle de la totalité des RDL, RCT et DDS (première table, lignes 1-5, plus ligne des totaux);
   - les deux catégories *v tr* et *v tr & v pron* se distinguent des autres par leur simple grandeur numérique (tables 1 et 3);
   - le fait qu'un verbe appartient à plusieurs catégories grammaticales, n'augmente pas forcément le nombre des indications sémantiques; la seule exception, c'est l'ensemble des *v tr & v pron*.

MIKLOS PÁLFY

Szeged