Banques de données textuelles, régionalismes de fréquence et régionalismes négatifs

André Thibault

André Thibault

Banques de données textuelles, régionalismes de fréquence et régionalismes négatifs*

1. Introduction

Le dépouillement manuel des sources, dans le domaine de la lexicologie différentielle, permet de mettre sur fiches des phénomènes ponctuels qui, en raison de leur caractère saillant, ont capté l'attention du chercheur. Quand il s'agit simplement d'attester l'existence d'un emploi lexical donné, quelques fiches de ce genre font l'affaire; en revanche, la méthode des dépouillements manuels traditionnels ne permet pas de rendre compte scientifiquement de l'existence de régionalismes de fréquence (c'est-à-dire d'unités lexicales plus fréquentes ou au contraire moins répandues dans une variété diatopique donnée que dans le reste de la francophonie), ainsi que de régionalismes négatifs (c'est-à-dire se signalant par leur quasi-absence dans une variété diatopique donnée; cette dernière catégorie n'étant que le cas extrême des régionalismes se signalant par une fréquence remarquablement basse).

Jusqu'à l'avènement des banques de données textuelles, les chercheurs avaient dû se contenter de remarques impressionnistes et subjectives sur la fréquence remarquablement basse ou élevée d'un emploi. Nous disposons aujourd'hui, et ce depuis quelques années déjà, de puissants outils informatiques et d'innombrables corpus de textes informatisés, qui nous permettent en théorie d'aborder de façon méthodologiquement renouvelée le problème des régionalismes négatifs et de fréquence. On constate toutefois que la recherche en lexicologie différentielle francophone, malgré le saut qualitatif auquel nous avons pu assister ces dernières années, n'a guère eu recours jusqu'à maintenant à cette nouvelle approche. Cela s'explique assez facilement lorsqu'on se penche sur la nature des outils existants, et surtout sur les nombreux problèmes que leur exploitation soulève. La route menant à une exploitation optimale des bases de données textuelles est semée d'embûches. C'est ce que nous allons tenter d'illustrer par quelques analyses de paires de mots, telles qu'on peut les étudier dans divers types de sources (dictionnaires conventionnels; dictionnaires de fréquence; bases de données textuelles).

* Nos remerciements s'adressent à Nathalie Bacon et Jean-François Smith, du Trésor de la Langue Française au Québec (Université Laval, Québec); à Christel Nissile, du Centre de dialectologie de l'Université de Neuchâtel (Suisse); à Geneviève Geron et Régine Wilmet, du Centre Valibel de l'Université de Louvain-la-Neuve (Belgique); enfin, à Robert Vézina, de l'Office Québécois de la Langue Française (Québec). Que tous veuillent bien trouver ici l'expression de notre gratitude pour l'aide qu'ils nous ont apportée dans l'exploitation des banques de données textuelles.

2. Un prédédent: l'exemple de *yogourt* vs *yaourt*

On trouve une première tentative, globalement satisfaisante, de ce que pourrait être un commentaire historico-comparatif différentiel francophone tenant compte de considérations quantitatives dans le DSR 1997 à l'article *yogourt*, mais cette tentative est restée isolée; ni les éditions successives du DSR (1999 et 2004), ni le DRF de Pierre Rézeau, ni le DHFQ de Claude Poirier n'ont explicitement cité de données quantitatives pour appuyer une affirmation portant sur la fréquence relative d'un emploi (alors que des régionalismes de fréquence figurent bel et bien dans ces ouvrages; cf. par ex. *arachide* dans le DHFQ, où le mot est présenté comme rare en France sur la base de sa représentation lexicographique lacunaire et sur les restrictions diaphasiques qu'il y connaît; cf. encore *couillon* dans le DRF, où la fréquence d'emploi est déduite de commentaires métalexicaux). L'utilisation explicite de données statistiques n'est pas à l'ordre du jour non plus dans le TLF; on y trouve bien, pour certains mots seulement, des indications de fréquence, mais elles ne sont jamais exploitées à des fins différentielles dans le cadre d'une analyse de nature diatopique. Encore eût-il fallu, en outre, que les rédacteurs aient été conscients du caractère éventuellement régional de certains phénomènes; le TLF donne par exemple comme première citation à l'article *brun* un passage d'une romancière canadienne (G. Guèvremont) en présentant comme ‹vx ou litt.› ce qui est en fait régional (il s'agit de la loc. verb. *faire brun* pour «s'assombrir (en parlant du ciel du crépuscule)», mal glosé par «sombre, obscur (en parlant de l'obscurité de la nuit)»).

Le cas de *yogourt* et *yaourt* se prêtait particulièrement bien à une recherche de nature statistique; ce lexème ne connaît guère qu'une seule acception, n'entre que dans très peu de locutions figurées, chacune des deux formes renvoie exactement au même référent, et il s'agit d'une lexie simple. Ce ne sont pas tous les couples de géosynonymes qui se trouvent dans la même situation. Nous allons examiner les cas suivants: *soulier / chaussure; brun / marron; orteil / doigt de pied.* Chacun d'entre eux présente des problèmes particuliers, qui compliquent l'étude de leur fréquence relative. Ils n'ont pas été traités jusqu'à présent dans le DSR, le DHFQ ou le DRF.

3. Le cas de *soulier* vs *chaussure*

L'expérience montre que Québécois et Français n'utilisent pas ces deux mots de la même manière. Devant une paire de chaussures prototypiques, le Québécois parlera spontanément de *souliers* là où le Français (à l'exception des locuteurs issus de régions archaïsantes comme par ex. le Bourbonnais; comm. de France Lagueunière) préférera tout aussi spontanément parler de *chaussures.* En franco-québécois, ce dernier mot fonctionne encore le plus souvent comme hypéronyme, alors que dans l'usage de l'Hexagone il a subi une restriction de sens et désigne dans la plupart des cas le plus prototypique de ses représentants, à savoir la chaussure basse, fermée, à semelle rigide, justement appelée *soulier* autrefois en France, et aujourd'hui encore au Québec. Il semble s'agir d'une assez banale situation d'innovation centrale et d'archaïsme périphérique, peut-être partagé par la

Suisse[1] et la Belgique. Voyons d'abord comment les dictionnaires conventionnels traitent la question.

3.1. Dictionnaires conventionnels

a) Petit Robert 2002. – Ce dictionnaire présente comme ‹vx ou région.› l'emploi de *soulier* au sens de «chaussure», mais néglige de préciser s'il s'agit de *chaussure* au sens 1 ou au sens 2, ce qui fait toute la différence. Il aurait fallu préciser ‹‹au sens 2››, mais alors on ne comprend plus quelle est la différence entre l'emploi français et l'emploi québécois du mot *soulier,* puisque la définition de ce mot est pratiquement identique à celle de *chaussure* dans son acception la plus courante (la seconde).

> **SOULIER** […] n. m. […] Chaussure à semelle résistante, qui couvre le pied sans monter beaucoup plus haut que la cheville. ⇒ **chaussure; brodequin, richelieu;** FAM. **croquenot, godasse,** 2. **pompe.** *Souliers bas, montants. Souliers plats. De vieux souliers. De gros souliers.* ◊ VX ou RÉGION. (Canada) Chaussure. *Souliers* (de femme) *à talons hauts. Souliers de marche, habillés. Souliers vernis. «Le Soulier de satin»,* drame de P. Claudel. […]
> CHAUSSURE […] n. f. […] 1. RARE (sens large) Partie du vêtement qui entoure et protège le pied. ⇒ babouche, 2. botte, bottillon, chausson, cothurne, espadrille, galoche, 2. mule, pantoufle, patin, sabot, sandale, savate, socque. […] 2. COUR. Chacun des deux objets fabriqués protégeant le pied, à semelle résistante, et qui couvre le pied sans monter plus haut que la cheville. ⇒ soulier; FAM. croquenot, godasse, godillot, grolle, 2. pompe, tatane. *Chaussure montante.* ⇒ boots, bottine, brodequin, 2. ranger. […] *Chaussures de ski.* […]

b) Hachette 1987 et DFPlus 1988. – Le second de ces dictionnaires est une adaptation québécoise du premier, mais n'a rien changé au texte; le passage ‹Syn. cour. de *soulier*› aurait dû disparaître dans le DFPlus.

> **soulier** […] n. m. Chaussure solide, à semelle rigide, couvrant le pied et, éventuellement, la cheville. *De gros souliers de marche.* > (Avec un qualificatif ou un comp.) Chaussure légère. *Des souliers vernis. Des souliers de daim.* […]
> **chaussure** […] n. f. **1.** Partie de l'habillement qui sert à couvrir et à protéger le pied (sandales, souliers, pantoufles, bottes, etc.); Syn. cour. de *soulier. Cirer, nettoyer, décrotter ses chaussures. Lacets, talons, semelles de chaussures.* […]. **2.** L'industrie de la chaussure.

c) DictUnivFr 1997 (autre adaptation du Hachette). – Ce dictionnaire panfrancophone présente comme québécisme le fait d'utiliser le mot *soulier* avec le sens de «chaussure» (entendu comme générique). Cet emploi n'est pas entièrement impossible en franco-québécois, mais le fait qu'on utilise très rarement *chaussure* comme hyponyme (et, par ricochet, très fréquemment *soulier* dans cette fonction) nous semble beaucoup plus important et n'apparaît pas dans les articles.

---

[1] «***soulier(s)*** n. m. (pl.) (*chaussures* est seulement générique en Suisse romande et au Québec et *soulier* s'y maintient comme le terme le plus courant). La restriction sémantique que connaît souvent le mot *chaussures* dans l'emploi hexagonal (où on l'emploie volontiers comme synonyme de *soulier* et pas seulement comme générique) est moins fréquente en Suisse romande et au Québec, où *soulier* reste encore le terme le plus courant pour désigner des chaussures rigides couvrant le pied mais non la cheville. Il faut donc classer l'emploi hexagonal comme une innovation sémantique.» Thibault 1996: 360.

soulier […] n. m. Chaussure solide, à semelle rigide, couvrant le pied et, éventuellement, la cheville. *De gros souliers de marche.* > (Avec un qualificatif ou un comp.) Chaussure légère. *Des souliers vernis. Des souliers de daim.* – (Québec) Cour. Chaussure (sens 1). […]
chaussure […] n. f. **1**. Partie de l'habillement qui sert à couvrir et à protéger le pied (sandales, souliers, pantoufles bottes, etc.). *Cirer, décrotter ses chaussures.* Syn. *soulier.* […] **2.** Industrie de la chaussure.

d) Adaptations du Micro-Robert. – Le Micro-Robert a donné lieu à deux adaptations, l'une pour le marché français et l'autre pour le marché québécois. Le premier précise avec beaucoup d'à-propos, s.v. *soulier,* que «dans l'usage courant, on dit *chaussure*»; la québécisation du DQA étant beaucoup plus approndie que celle du DFPlus, cette remarque n'y apparaît pas. En revanche, la seconde acception de *chaussure,* l'hyponymique, apparaît sans aucun commentaire dans le DQA (ce qui n'est pas entièrement injustifié, dans la mesure où cet usage tend à se répandre au Québec, en particulier dans la langue commerciale).

RobAuj 1991: ***soulier*** […] Chaussure à semelle résistante, qui couvre le pied sans monter beaucoup plus haut que la cheville. ≠ *chausson, botte, bottine. Souliers de marche, habillés, de sport.* – REM. Dans l'usage courant, on dit *chaussure,* sauf en parlant des gros *souliers* de marche. […] – ***chaussure*** […] **1.** Partie du vêtement qui protège le pied. *Des gens qui marchent sans chaussures.* **2.** Chaussure (1) solide, basse et fermée (opposé à *chausson, sabot, sandale, botte*). ⇒ **soulier;** fam. **godasse, grole, pompe, tatane**. *Chaussure de marche, de sport.* ⇒ **mocassin;** 2. **basket, tennis.** *Chaussures habillées.* ⇒ **escarpin.** *Faire réparer des chaussures chez le cordonnier.* […]
DQA 1992: ***soulier*** […] Chaussure à semelle résistante, qui couvre le pied sans monter beaucoup plus haut que la cheville. ⇒ **mocassin;** fam. **pichou.** ≠ *botte, bottine, chausson, pantoufle, sandale. Souliers de marche, habillés, de sport.* […] – ***chaussure*** […] **1.** Partie du vêtement qui protège le pied. *Des gens qui marchent sans chaussures.* **2.** Chaussure (1) solide, basse et fermée (opposé à *chausson, sabot, sandale, botte*). ⇒ **soulier;** fam. **godasse**. *Chaussure de marche, de sport.* ⇒ **mocassin;** 2. **basket, espadrille, tennis.** *Chaussures habillées.* ⇒ **escarpin.** *Faire réparer des chaussures chez le cordonnier.* […]

e) DQF 1999. – Cet ouvrage, qui s'est donné pour but de relever toutes les différences entre le franco-québécois et le français de France, signale la fréquence particulière du mot *chaussure* en France au sens hyponymique:

soulier […] – ***acheter une paire de souliers:*** une paire de chaussures. Auj., en français standard [*sic*], le mot le plus général est «chaussure»; le mot «soulier» est considéré comme vieil., sauf s'il désigne un type particulier de chaussures, surtout des chaussures solides ou prévues pour un usage part.: «souliers de marche», «gros souliers», «souliers ferrés», etc. […].

f) RobHist 1992. – Ce dictionnaire signale que le mot *chaussure* domine, non seulement comme hyponyme (s.v. *soulier*), mais aussi comme hypéronyme (s.v. *chaussure*); l'aspect diachronique est effleuré et la variation diatopique, ignorée.

SOULIER […] Dans l'usage courant, on emploie aujourd'hui plus souvent *chaussure.*
CHAUSSURE […] Malgré la concurrence de *soulier* et de termes spécifiques, *chaussure* est resté le mot générique usuel […].

g) TLF. – On trouve dans le TLF des indications de fréquence, absolue et relative, mais pas pour tous les mots. Ces données sont toutefois assez dépourvues d'intérêt. Comme la base Frantext a évolué au fur et à mesure de la rédaction du TLF, les comparaisons entre articles ne valent pas grand-chose. En outre, la polysémie n'est pas prise en charge par ces statistiques brutes (par ex., *marron* n. m. désignant le fruit, et *marron* adj. de couleur), ni les lexies complexes (comme *petit déjeuner* ou *doigt de pied*). Les résultats pour *chaussure* et *soulier* sont toutefois très indicatifs, mais ne portent que sur le 19e et le 20e siècle, alors

que Frantext inclut des textes du 16e siècle à nos jours. Nous avons donc refait les statistiques directement à partir de cette base (v. ci-dessous 3.3.c).

3.2. Dictionnaires de fréquence (Juilland 1970 et DFréQ 1992)

On pourrait croire, a priori, que les dictionnaires de fréquence sont les instruments par excellence pour étudier le problème qui nous occupe. Or, rien n'est plus faux. Nous ne disposons de dictionnaires de fréquence, en francophonie, que pour la France et le Québec. À vrai dire, Juilland 1970 n'est pas parfaitement homogène du point de vue diatopique (on y trouve quelques auteurs non français, comme le Belge Crommelynck), mais on peut tout de même le considérer comme largement représentatif du français de France. Quant au DFréQ 1992, il est homogène du point de vue diatopique mais en toute rigueur on ne peut le comparer au Juilland, car une bonne partie de son corpus est tirée de la langue orale spontanée (alors que le Juilland est aussi littéraire que peut l'être le TLF). D'entrée de jeu, on constate que les données ne sont pas réellement comparables, en raison de la différence diamésique qui sépare les deux ouvrages. Le second problème n'est pas moins grave: sans vouloir jeter la première pierre, car nous savons que la lemmatisation de masses de matériaux textuels demande beaucoup d'interventions manuelles et semi-automatiques, les corpus ayant donné lieu à l'élaboration de ces dictionnaires sont largement insuffisants (500.000 mots pour le Juilland, un million de mots pour le DFréQ). Les résultats paraissent donc largement aléatoires. Le mot *chaussure* est tout simplement absent du Juilland; quand au mot *soulier,* il y apparaît avec une fréquence absolue de 7, un taux de dispersion de 46,54 et un taux d'usage de 3,25 qui le met au 4836^{e} rang d'une liste de 5083 mots. L'absence du mot *chaussure* dans Juilland n'est qu'un exemple de plus de la piètre représentativité de ce genre d'instruments. Paradoxalement, *chaussure* apparaît 12 fois dans le DFréQ, avec un indice de dispersion de 59,17 et un indice d'usage de 7,1; on pourrait donc en conclure qu'il est plus fréquent au Québec qu'en France! Le mot *soulier* est toutefois beaucoup mieux représenté, avec une fréquence absolue de 42, un taux de dispersion de 67,51 et un taux d'usage de 28,3 (donc beaucoup plus élevé que *chaussure*).

3.3. Bases de données textuelles

a) Québétext. – Au 19^{e} siècle, on relève 50 att. de *soulier(s)* pour 11 att. de *chaussure(s)*, pour un rapport de 82,0/18,0; au 20^{e} siècle, 310 att. de *soulier(s)* pour 131 att. de *chaussure(s)*, pour un rapport de 70,3/29,7. Ces résultats peuvent s'interpréter de deux façons: d'un point de vue diatopique, on constate qu'en franco-québécois (à tout le moins dans la langue littéraire) le mot *soulier* domine largement le mot *chaussure*, ce qui semble être le contraire de ce que l'on observe en France; d'un point de vue diachronique, le mot *chaussure* semble tout de même avoir fait des gains, l'écart entre les deux s'étant amenuisé d'un siècle à l'autre. L'innovation centrale atteint tout de même peu à peu la périphérie.

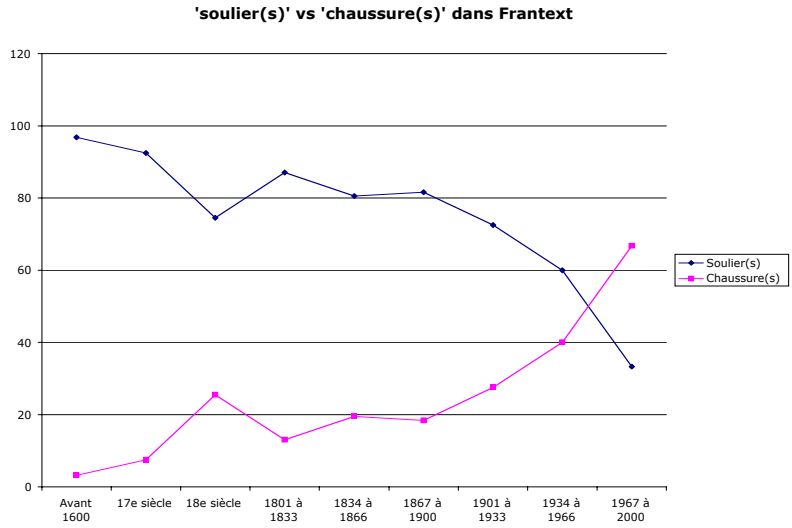
b) Suistext. – 323 att. de *soulier(s)* pour 162 att. de *chaussure(s),* pour un rapport de 66,7/33,3. Rappelons que Suistext ne réunit que des écrivains romands contemporains (la

plupart sont encore vivants aujourd'hui). On voit qu'ici aussi, *soulier* est beaucoup plus fréquent que *chaussure,* ce qui semble confirmer que la Suisse romande fonctionne bien comme une périphérie conservatrice (bien qu'elle soit beaucoup moins éloignée que le Québec) – à tout le moins pour ce mot (le français de Suisse romande produit aussi de nombreuses innovations). Des observations diachroniques ne sont malheureusement pas possibles; on constate seulement que le rapport de deux tiers / un tiers se rapproche davantage du rapport observé chez les textes québécois du 20^{e} siècle que du 19^{e}.

c) Frantext. – Première question: Frantext est-il un outil fiable pour l'étude des variations de fréquence sur l'axe diatopique? La réponse est négative. Il n'existe aucun filtre automatique dans Frantext qui permette de créer un sous-ensemble de textes regroupant les auteurs selon leur origine géographique. Il n'est malheureusement pas possible, à moins de le faire manuellement, de regrouper, par exemple, tous les ouvrages rédigés par des Canadiens, des Belges ou des Suisses, encore moins par des Provençaux. Il n'est pas non plus possible de créer automatiquement un sous-ensemble de textes ne regroupant que les auteurs de l'Hexagone. Autant les études diachroniques sont faciles à mener dans Frantext, autant les interrogations diatopiques sont rendues impossibles en raison des lacunes conceptuelles qui caractérisent cet instrument. Lorsqu'il s'agit de traiter de petites quantités d'occurrences, on peut toujours aborder les matériaux manuellement; c'est ce que nous avons fait pour *yogourt / yaourt,* en mettant de côté les exemples du Suisse Benoziglio. En revanche, lorsque la masse des matériaux dépasse les quelques centaines d'attestation, Frantext devient ingérable du point de vue diatopique. Il faut donc souhaiter que les responsables de cet outil informatique incomparable pensent à ajouter cette fonction à leur liste de possibilités de constitution de «corpus». D'ici là, nous en sommes réduits à accepter un certain flou dans les résultats, dont l'importance n'est pas vraiment si grande dans la mesure où Frantext n'accueille qu'un tout petit pourcentage d'auteurs non français.

Cela dit, Frantext donne des résultats très intéressants pour l'étude diachronique des relations entre *soulier* et *chaussure,* dont on trouvera le détail ci-dessous, et leur représentation graphique à la suite.

Période	soulier(s)	chaussure(s)	rapport	Période	soulier(s)	chaussure(s)	rapport
Av. 1600:	92	3	96,8/3,2	1867-1900:	759	171	81,6/18,4
1601-1700:	197	16	92,5/7,5	1901-1933:	817	310	72,5/27,5
1701-1800:	313	107	74,5/25,5	1934-1966:	1147	764	60,0/40,0
1801-1833:	181	27	87,0/13,0	1967-2004:	415	830	33,3/66,7
1834-1866:	908	220	80,5/19,5				



d) Les cédéroms de la presse quotidienne et hebdomadaire. – Les cédéroms réunissant le texte intégral de différents quotidiens et hebdomadaires, en France comme au Québec, représentent une véritable mine d'or pour le lexicographe qui recherche des formes rares. Leur rendement en ce qui concerne les régionalismes sémantiques est déjà beaucoup plus limité, en raison de l'énorme «bruit» généré par les interrogations. Lorsqu'il s'agit, comme ici, de comparer les proportions respectives de paires de lexèmes de part et d'autre de l'Atlantique, les problèmes varient beaucoup selon les cas. Comme nous l'avons déjà signalé à propos de l'article *yogourt* du DSR, des lexies simples, relativement rares et monosémiques se prêtent particulièrement bien à des recherches dans ce genre de support. Le couple *soulier / chaussure* permet d'arriver à des résultats probants. Cela dit, les cédéroms de presse disposent de moteurs de recherche tout à fait inappropriés par rapport au type d'interrogations que nous souhaiterions mener. Ils n'offrent même pas de concordances, mais simplement la liste des articles dans lesquels figure le mot recherché; il faut lire chaque article pour étudier manuellement la syntagmatique, l'environnement syntaxique, etc. Ils ne permettent pas de construire des interrogations à la syntaxe élaborée et sont très en deçà, de ce point de vue, de bases comme Frantext, Québétext ou Suistext dans le monde francophone, ou de CORDE et CREA dans le monde hispanophone (v. www.rae.es); quand on sait que, pour les spécialistes, ces bases faites pour une exploitation linguistique sont déjà considérées comme primitives, on conviendra que les moteurs de recherche des cédéroms de presse sont de véritables dinosaures, totalement inadaptés à nos besoins.

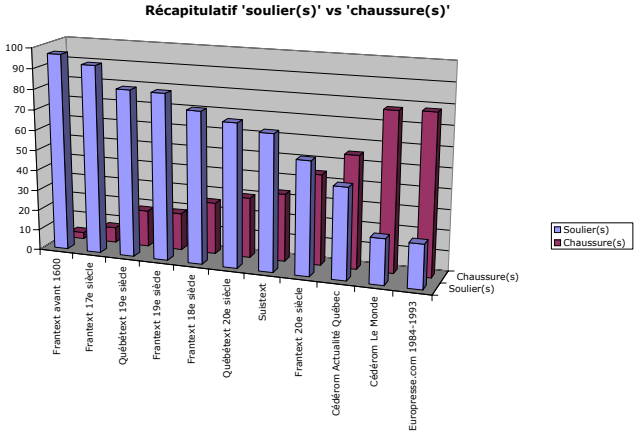
Le cédérom du journal *Le Monde,* texte intégral depuis 1987. Instrument bien connu déjà des lexicologues, qui l'utilisent avec profit pour repérer des mots rares. Le DRF a beaucoup enrichi sa récolte de citations en l'exploitant. Résultats: *soulier(s)*: 1192; *chaussure(s)*: 4156; rapport: 22,3/77,7.

Le cédérom Europresse.com. Cette banque de données textuelles réunit le texte intégral d'un grand nombre de publications françaises, dont la plus ancienne remonte à 1984. Son utilisation n'est guère conviviale dans l'optique du linguiste (pas de concordances, pas d'opérateurs booléens), et les résultats sont faussés par le fait qu'une même dépêche peut se répéter dans plusieurs journaux. Ce défaut peut être corrigé manuellement

lorsqu'il ne s'agit que d'un petit nombre d'attestations, mais tout l'intérêt de cette base est qu'elle est gigantesque et fournit des masses de matériaux considérables. Idéalement, une utilisation de cette base à des fins statistiques devrait pouvoir régler automatiquement ce problème de redondance. Un autre problème est constitué par le fait que le programme ne fournit que les 1000 premiers documents trouvés. Dans le cas de mots très fréquents, comme *chaussure(s),* il aurait pratiquement fallu scinder le corpus en tranches de six mois, ce qui sur une période de vingt ans devient vite assez laborieux et chronophage. Ce problème devrait aussi être réglé, dans l'optique d'une exploitation statistique efficace. Voici néanmoins les résultats obtenus pour la tranche chronologique 1984 à 1993: *soulier(s)*: 665; *chaussure(s)*: 2402; rapport: 21,7/78,3.
Le cédérom Actualité Québec. Sondages portant sur les années 1985, 1990, 1995, 2000, 2003 et 2004:*soulier(s)*: 2196; *chaussure(s)*: 2715 doc.; rapport: 44,7/55,3. Contrairement aux données de Québétext, qui représentent la langue littéraire, la langue de la presse donne la priorité à *chaussure;* toutefois, en comparaison avec *Le Monde,* on voit que cette priorité est beaucoup plus faible. Il convient ici d'ajouter une importante remarque de nature méthodologique: les journaux québécois ne représentent pas des objets ‹‹purs›› du point de vue diatopique. On y trouve en effet de nombreuses dépêches d'agence de presse internationales, ainsi que des articles tirés de grands quotidiens français tels *Le Monde* ou *Libération* qui paraissent régulièrement dans les colonnes de certains journaux québécois avec lesquels ils semblent avoir passé des accords de coopération. Cela peut expliquer partiellement la domination statistique de *chaussure* sur *soulier.* Il n'existe malheureusement aucune façon automatique de mettre de côté les articles qui n'ont pas été rédigés par des journalistes québécois; il faudrait, idéalement, faire le tri manuellement, ce qui devient vite très laborieux lorsqu'on a affaire à de telles masses de matériaux. Une utilisation méthodologiquement irréprochable de cet outil impliquerait que l'on puisse séparer les articles selon la provenance de leur rédacteur. On a d'ailleurs déjà pu dire la même chose de Frantext (voir ci-dessus).

Récapitulatif *soulier(s)* vs *chaussure(s)*

Frantext avant 1600:	96,8/3,2	Suistext:	66,6/33,3
Frantext 17e siècle:	92,5/7,5	Frantext 20e siècle:	55,5/44,5
Québétext 19e siècle:	82,0/18,0	CD-rom *Actualité Québec:*	44,7/55,3
Frantext 19e siècle:	81,6/18,4	CD-rom *Le Monde* (dp. 1987):	22,3/77,7
Frantext 18e siècle:	74,5/25,5	Europresse.com 1984-1993:	21,7/78,3
Québétext 20e siècle:	70,3/29,7		



Les proportions respectives de ces deux mots ont connu une évolution très rapide au cours du siècle dernier. Il est intéressant de voir que la base Québétext illustre un usage ayant eu cours en France jusqu'au 19e siècle; Suistext est légèrement moins archaïque, mais n'atteint pas les chiffres étonnants relevés dans le cédérom *Actualité Québec,* qui montrent un alignement croissant de la langue de la presse québécoise sur l'usage de France, sans

toutefois aller jusqu'aux extrêmes enregistrés dans les plus récents cédéroms de la presse française. Ces résultats montrent avec éloquence qu'il est possible, grâce aux banques de données textuelles pan-francophones, de tendre un pont entre diachronie et diatopie.

4. Le cas de *brun* vs *marron*

Pour un locuteur québécois ou suisse romand, *marron* comme adjectif de couleur est plutôt inusité et senti comme un emprunt occasionnel au français de France. C'est un cas typique, à notre sens, de régionalisme négatif. Il est toutefois exceptionnel que la lexicographie rende compte de cette réalité.

4.1. Dictionnaires conventionnels

a) DQA 1992. – Ce dictionnaire qui épingle les ‹‹francismes›› attire l'attention du lecteur québécois sur le fait qu'en France, *marron* s'emploie comme adjectif de couleur invariable; il l'exemplifie de manière plausible avec *des robes marrons.*

> ***marron*** […] n. m. et adj. inv. […] **3.** (France) Adj. invar. D'une couleur brune et foncée. *Des robes marrons.*

b) DictUnivFr 1997. – Ce dictionnaire panfrancophone laisse deviner au lecteur, sans l'affirmer explicitement, que la particularité diatopique québécoise consiste à employer l'adjectif *brun* avec un nom de vêtement (ce qui, on l'a vu ci-dessus dans le DQA, est plutôt la prérogative de *marron* en français de France).

> **brun, brune** […] (Québec) Un chandail brun pâle, clair, foncé.

c) DQF 1999. – Quant à cet ouvrage, il suggère (sans le dire explicitement lui non plus) que *brun* cède le pas à *marron* dans l'usage de France pour désigner des vêtements (manteau, chaussures; ‹porter›). Il déclare en outre un peu imprudemment que *marron* est inusité au Québec, ce qui n'est pas entièrement vrai pour la langue écrite (on trouve sans difficultés des dizaines d'attestation du mot dans la presse sur cédérom), mais correspond bien à notre sentiment en ce qui concerne la langue orale.

> **brun** […] – **manteau brun:** manteau marron / – **souliers bruns:** chaussures marron / – **porter du brun:** porter du marron / En français standard [*sic*], pour caractériser un objet de couleur brun-rouge, on dit cour. «marron» […]; de ce fait, la fréq. d'emploi de l'adj. «brun» est beaucoup moins élevée; au Québec, on n'utilise pas l'adj. «marron».

d) TLF. – Le principal point fort de l'article *brun* du TLF est qu'il nous indique quelles sont les préférences colocationnelles de cet adjectif: *ours, cheveux, châtaignes, chemises* (mais il s'agit là d'une lexie figée), *bras, labours, cigarette, bière, teint*. À aucun moment on ne fait référence à la répartition distributionnelle entre *brun* et *marron.* Celui-ci, en tant qu'adjectif de couleur, n'a droit qu'à une subdivision de l'article consacré au substantif désignant le fruit. On y apprend que *marron* peut apparaître aux côtés des mots *caban*, *selle*, *costume*, *redingote*, *pardessus*, *drap*, *lainage*, *soie*, et *yeux*. Encore une fois, aucune

allusion n'est faite à *brun* et aux rapports que *marron* entretient avec ce dernier. C'est en fait d'une façon presque involontaire que les articles du TLF nous sont utiles ici: c'est la distribution de chaque adjectif, telle qu'on peut l'observer dans les collocations, les exemples enchaînés et les exemples cités, qui nous permet d'entrevoir comment ceux-ci se répartissent les référents. L'adjectif *brun* semble dominer avec les référents «organiques», alors que *marron* semble s'être spécialisé dans les référents «non-organiques», en particulier des étoffes ou des vêtements; mais nous verrons que le substantif *yeux* échappe à cette répartition.

4.2. Dictionnaires de fréquence

Quant aux dictionnaires de fréquence, on ne peut que constater leur complète inutilité en ce qui concerne des mots polysémiques comme *brun* et *marron;* ce dernier, en particulier, correspond à trois entrées séparées dans la plupart des dictionnaires, mais l'on sait que les dictionnaires de fréquence ne pratiquent qu'exceptionnellement le dégroupement des homonymes (cf. *défendre* «prendre la défense de» et «interdire» dans Juilland), et pratiquement jamais celui des acceptions d'un lexème polysémique. De toute façon, *marron* est complètement absent du Juilland (qui donne toutefois 17 attestations de *brun(e)(s),* toujours adjectif), ainsi que du DFréQ (qui donne quant à lui 13 attestations de *brun(e)(s),* 12 fois adj. et une fois subst.). Si l'on ne se fiait qu'à ces dictionnaires, il faudrait en conclure que *brun* est plus fréquent que *marron* de toute façon, autant en France qu'au Québec – ce qui n'est d'ailleurs pas impossible.

4.3. Bases de données textuelles

La quantité ingérable d'attestations de *brun* et de *marron,* en particulier à cause de la polysémie de ce dernier, rendait impossible une comparaison entre ces deux adjectifs. Nous avons donc choisi d'étudier un couple qui nous semblait intéressant car révélateur de phénomènes diachroniques et diatopiques: *yeux bruns* vs *yeux marron(s).*

Suistext: *yeux bruns*: 21; *yeux marron(s)*: 5 (tous chez C. Bille); rapport: 80,8/19,2. Ici, la dispersion du second est très mauvaise, tous les exemples apparaissant chez la même romancière. Son score relativement élevé n'est donc pas très significatif.

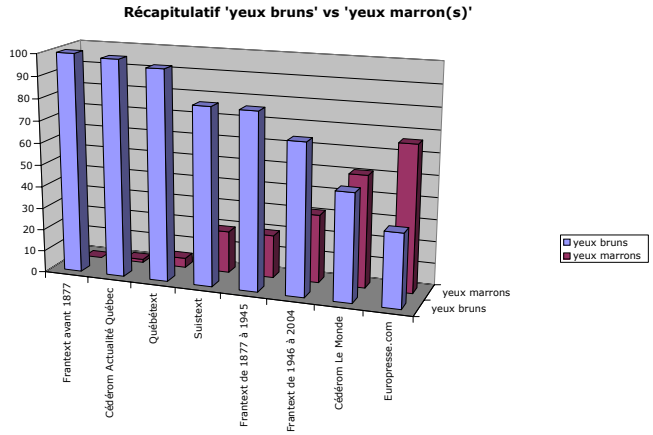
Québétext: *yeux bruns*: 22; *yeux marron*: 1; rapport: 95,7/4,3.
Frantext: Rapport *yeux bruns / yeux marron(s)* avant 1877: 100/0 (*yeux marron(s)* n'est pas attesté); rapport de 1877 à 1945: 90 pour 22, donc 80,4/19,6; rapport de 1946 à nos jours: 44 pour 20, donc 68,8/31,2.

Les cédéroms de presse quotidienne et hebdomadaire:

a) Le cédérom du journal *Le Monde,* texte intégral depuis 1987: *yeux bruns*: att. dans 23 doc.; *yeux marron(s)*: att. dans 24 doc.; rapport: 48,9/51,1.
b) Europresse.com: *yeux bruns*: att. dans 105 doc.; *yeux marrons*: att. dans 208 doc.; rapport: 33,5/66,5.
c) Le cédérom *Actualité Québec*: sondages portant sur les années 1985, 1990, 1995, 2000, 2003 et 2004: *yeux bruns*: att. dans 217 doc.; *yeux marron(s)*: att. dans 4 doc. (dont un article rédigé par un journaliste d'origine française); rapport: 98,6/1,4.

4.4. Récapitulatif pour *yeux bruns* vs *yeux marron(s)*

Frantext avant 1877:	100/0	Frantext de 1877 à 1945:	80,4/19,6
Cédérom *Actualité Québec:*	98,6/1,4	Frantext de 1946 à nos jours:	68,8/31,2
Québétext:	95,7/4,3	Cédérom *Le Monde:*	48,9/51,1
Suistext:	80,8/19,2	Europresse.com:	33,5/66,5



Le syntagme *yeux marrons* est d'apparition très récente dans la langue française (1877 dans Frantext); à vrai dire, l'emploi de *marron* comme adjectif de couleur est lui aussi relativement récent (1706, *couleur de maron;* 1750 *maron* comme nom de couleur; v. TLF). Contrairement à ce que nous avons pu observer pour le couple *soulier / chaussure,* les cédéroms de la presse québécoise ne permettent pas d'observer une propagation de l'innovation française, qui ne se glisse encore que très timidement dans Québétext (peut-être dans des textes d'agences de presse françaises), et plus nettement dans Suistext (mais avec une dispersion nulle qui empêche d'en tirer des conclusions significatives). Dans Frantext, on assiste à une forte évolution depuis la fin de la Seconde Guerre Mondiale, encore accentuée dans la dernière décennie dans les cédéroms de la presse française. Il semble bien qu'en France, *yeux bruns* soit en train de céder la place à *yeux marron(s),* en dépit de la répartition «organique / non-organique» qui régit les affinités syntaxiques des deux adjectifs.

5. Le cas de orteil vs doigt de pied

Nous savons par expérience qu'au Québec la lexie *doigt de pied* ne fait pas partie de la langue courante et est franchement perçue comme appartenant au français de France, son synonyme *orteil* étant le seul employé en franco-québécois oral spontané. En fait, *doigt(s) de pied* serait à l'origine un régionalisme nord-oriental, s'étant diffusé hors de sa sphère d'origine aussi tard qu'au 20^{e} siècle (v. Chambon 1991 et FEW 25, 381a, n. 32).

a) Dictionnaires conventionnels. – Aucun dictionnaire, toutefois, ne commente la nature de régionalisme négatif de cette lexie; ni le DQA de Jean-Claude Boulanger, pourtant si

prompt à taxer de «francisme» les mots et acceptions peu répandus au Québec, ni le DQF de Lionel Meney, toujours si désireux d'enseigner aux Québécois le «français standard» (il ne commente *orteil* que pour rappeler que ce mot est bien de genre masculin), ne mentionnent le phénomène.

b) Dictionnaires de fréquence. – Quant aux dictionnaires de fréquence, nous pouvons bien vite les écarter: aucun d'entre eux n'a traité comme une lexie à part entière le syntagme figé *doigt de pied.*

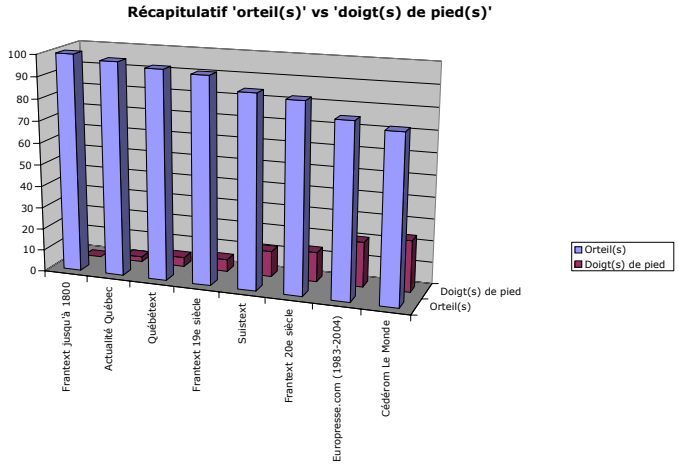
c) Bases de données textuelles. – Suistext: *orteil(s)*: 45; *doigt(s) de pied(s)*: 6; rapport: 88,2/11,8. – Québétext: *orteil(s)*: 109; *doigt(s) de pied(s)*: 10; rapport: 91,6/8,4. – Frantext: jusqu'à 1800: *orteil(s),* 46; *doigt(s) de pied(s),* 0; rapport: 100/0; 19e siècle: *orteil(s),* 169; *doigt(s) de pied(s),* 10; rapport: 94,4/5,6; 20e siècle: *orteil(s),* 452; *doigt(s) de pied(s),* 72; rapport: 86,3/13,7.

Les cédéroms de presse quotidienne et hebdomadaire. – Le cédérom du journal *Le Monde* (texte intégral depuis 1987): *orteil(s)*: att. dans 325 doc.; *doigt(s) de pied(s)*: att. dans 101 doc.; rapport: 76,3/23,7. Ce rapport est légèrement plus favorable à *doigt(s) de pied* que celui relevé dans Frantext tout au long du 20e siècle. – Europresse.com (totalité de la base, 1983-2004): *orteil(s):* att. dans 1921 doc.; *doigt(s) de pied(s)*: att. dans 497 doc.; rapport: 79,4/20,6. – Le cédérom *Actualité Québec*, sondages portant sur les années 1985, 1990, 1995, 2000, 2003 et 2004: *orteil(s):* att. dans 587 doc.; *doigt(s) de pied(s)*: att. dans 13 doc.; rapport: 97,8/2,2. Ici, en revanche, le rapport est franchement défavorable à *doigt(s) de pied(s);* dans la presse québécoise, *orteil(s)* domine très largement.

d) Récapitulatif pour *orteil(s)* vs. *doigt(s) de pied(s)*:

Frantext jusqu'à 1800:	100/0	Suistext:	88,2/11,8
Actualité Québec:	97,8/2,2	Frantext 20e siècle:	86,3/13,7
Québétext:	95,7/4,3	Europresse.com (1983-2004):	79,4/20,6
Frantext 19e siècle:	94,4/5,6	Cédérom *Le Monde* (dp. 1987):	76,3/23,7

Encore une fois, le cédérom de la presse québécoise ainsi que la base de textes littéraires Québétext nous révèlent un usage plus proche de celui du français classique. L'évolution semble s'être accélérée récemment en France, la langue de la presse contemporaine étant plus favorable à *doigt(s) de pied(s)* que le sous-ensemble de Frantext regroupant les textes du 20e siècle.



6. Conclusion

Quel bilan peut-on établir à l'issue de ce tour d'horizon? En ce qui concerne Frantext, on déplore l'impossibilité de faire des sous-ensembles de corpus diatopiquement cohérents; les types d'interrogations offerts par cette base sont toutefois plus satisfaisants que ceux des cédéroms de la presse, qui ne sont pas du tout adaptés aux besoins des linguistes (absence de concordances, pas de fonctions de recherche statistique, aspect trop limité des recherches de co-occurrences, difficulté de désambiguïser rapidement et efficacement les résultats obtenus, redondance des résultats dans Europresse.com). Quant à Québétext et Suistext, on regrettera leur taille encore relativement modeste, qui ne garantit pas toujours une représentativité satisfaisante, et un manque de profondeur historique, en particulier pour Suistext, qui ne permet pas la comparaison avec les données de Frantext (ce manque de profondeur historique est encore plus aigu dans les cédéroms de la presse francophone). De manière générale, on notera que la polysémie est un obstacle grave à l'exploitation intensive des bases de données textuelles. Quant aux aspects positifs, nous croyons avoir démontré à l'aide de quelques exemples qu'avec des paires de lexies raisonnablement monosémiques, les banques textuelles peuvent nous apporter des données objectives très précieuses pour l'étude de certains aspects diachroniques et diatopiques de l'évolution lexicale du français, qu'un travail de mise en fiches traditionnel ne saurait jamais remplacer.

Références bibliographiques

Belg 1994 = Bal, Willy et al. (1994): *Belgicismes. Inventaire des particularités lexicales du français en Belgique.* Louvain-la-Neuve: Duculot.

Chambon, Jean-Pierre (1991): À propos de gros sous et de doigts de pied chez Rimbaud. In: *Parade sauvage: revue d'études rimbaldiennes*, n° 8, 9-15.

DFPlus 1988 = Poirier, Claude (rédacteur principal) (1988): *Dictionnaire du français plus: à l'usage des francophones d'Amérique.* Montréal: Centre éducatif et culturel Inc.

DFréQ 1992 = Beauchemin, Normand, Pierre Martel, Michel Théoret (1992): *Dictionnaire de fréquence des mots du français parlé au Québec.* Berne: Peter Lang.

DHFQ 1998 = Poirier, Claude (dir.) (1998): *Dictionnaire historique du français québécois.* Sainte-Foy (Québec): Les Presses de l'Université Laval.

DictUnivFr 1997 = *Dictionnaire universel francophone.* Paris: Hachette/EDICEF/AUPELF-UREF.

DQA 1992 = Boulanger, J.-Cl. (1992): *Dictionnaire québécois d'aujourd'hui.* Montréal: Dicorobert.

DQF 1999 = Meney, Lionel (1999). *Dictionnaire québécois français.* Montréal: Guérin.

DRF = Rézeau, Pierre (dir.) (2001): *Dictionnaire des régionalismes de France. Géographie et histoire d'un patrimoine linguistique.* Bruxelles: De Boeck/Duculot.

DSR 1997 = Thibault, André (1997): *Dictionnaire suisse romand: particularités lexicales du français contemporain.* Genève-Carouge: Zoé.

DSR 1999 = Thibault, André (1997): *Dictionnaire suisse romand, éd. sur cédérom, revue et enrichie.* Genève-Carouge: Zoé.

DSR 2004 = Thibault, André (2004): *Dictionnaire suisse romand: particularités lexicales du français contemporain.* Nouvelle éd. rev. et augm. préparée par P. Knecht. Genève-Carouge: Zoé.

Europresse.com = Presse d'information francophone en texte intégral (réunit le texte des publ. suivantes: AFP Général, dp. 19/03/01; L'Express, dp. 7/01/93; L'Humanité, dp. 16/11/99; La Croix, dp. 1/09/95; Le Figaro, dp. 31/10/96; Le Monde, dp. 1/01/97; Le Monde Diplomatique, dp. 1/01/84; Le Parisien, dp. 2/05/98; Le Point, dp. 7/01/95; Les Échos, dp. 2/01/91; Libération dp. 2/01/95).

FEW = Wartburg, W. von (1922-2002): *Französisches Etymologisches Wörterbuch. Eine Darstellung des galloromanischen Sprachschatzes.* Bonn/Leipzig/Bâle: Teubner/Klopp/Zbinden, 25 vol.

Frantext = Base de données textuelles FRANTEXT, gérée par le logiciel STELLA, Nancy, Centre National de la Recherche Scientifique, ATILF (ex-INaLF). Consultable en ligne sur abonnement.

Hachette 1987 = *Le dictionnaire du français.* Paris: Hachette.

HanseChasse 1980: Hanse, J. et al. (1980): Chasse aux belgicismes. Bruxelles: Fond. Ch. Plisnier.

Juilland 1970 = Juilland, Alphonse, Dorothy Brodin, Catherine Davidovitch, with the collaboration of Mary Ann Ignatius, Ileana Juilland, Lilian Szklarczyk (1970): *Frequency dictionary of French words.* Paris / La Haye: Mouton.

Québétext = banque de données textuelles réalisée par le Trésor de la Langue Française au Québec (Université Laval, Québec), réunissant (entre autres) des textes littéraires d'auteurs québécois du 19^{e} et du 20^{e} siècles; les textes littéraires de 1837 à 1919 sont en accès libre à l'adresse suivante: www.tlfq.ulaval.ca/quebetext/.

RobAuj 1991 = *Le Robert dictionnaire d'aujourd'hui.* Paris: Dictionnaires Le Robert.

RobHist 1992 = Rey, A. (1992): *Dictionnaire historique de la langue française.* Paris, Le Robert.

Suistext = banque de données textuelles réalisée par la station suisse du Trésor des Vocabulaires francophones de l'Université de Neuchâtel, et contenant la totalité de l'œuvre de quatorze écrivains romands contemporains (É. Barilier, C. S. Bille, G. Borgeaud, N. Bouvier, M. Chappaz, J. Chessex, C. Colomb, A.-L. Grobéty, J.-M. Lovay, J. Mercanton, J.-P. Monnier, A. Rivaz, Y. Velan, A. Voisard). N'est consultable que sur place.

Thibault, André (1996): Québécismes et helvétismes: éclairages réciproques. In: Lavoie, Thomas (éd.) (1996): *Français du Canada – Français de France.* Tübingen: Niemeyer, 333-376.

TLF = C.N.R.S. (éd.) (1971-1994): *Trésor de la langue française: Dictionnaire de la langue du XIX^{e} et du XX^{e} siècles (1789-1960).* Paris: Gallimard, 16 vol.