Synthèse des avancées du projet PRESTO

Juin 2014

1. Corpus

1.1. Structure du corpus

La constitution du corpus PRESTO s'appuie sur la collaboration avec diverses bases textuelles présentes en France (BVH, FRANTEXT, …) et à l'étranger (Cologne, ARTFL). L'objectif est de construire un corpus à trois niveaux :

- **Un premier niveau**, ou ***corpus noyau***, composé d'un ensemble restreint de textes (échantillonnés s'ils dépassent 50 000 mots), diffusable librement sous licence Creative Commons (le type de licence est en cours de négociation avec les différents partenaires). Ces textes seront lemmatisés et annotés morpho-syntaxiquement puis vérifiés et corrigés manuellement.
- **Un deuxième niveau**, ou ***corpus second*****,** beaucoup plus vaste et probablement moins ouvert, contiendra les textes du corpus noyau (en version intégrale pour ceux qui avaient été échantillonnés), les œuvres complètes de certains auteurs du corpus noyau (dans la mesure du possible) et d'autres textes permettant d'étoffer et de diversifier les genres textuels du corpus noyau. Un travail important a d'ores et déjà accompli à Cologne pour constituer un corpus de presse diachronique (cf. *infra*) versé dans ce deuxième niveau. L'intégalité du corpus second sera lemmatisée et annotée morpho-syntaxiquement mais non vérifiée manuellement.
- **Un troisième niveau**, moins structuré, permettra d'inclure des textes provenant d'origines diverses, au gré des accords noués avec divers collègues, mais ne répondant pas à des critères de sélection aussi stricts (si ce n'est concernant leur qualité). Cela permettra de diversifier plus encore les genres et les domaines mais également les types de textes (intégration de manuscrits par exemple) et d'accroître le volume textuel.

1.2. Corpus noyau : critères et état d'avancement

Une première version du corpus noyau a été élaborée pour la période XVI^{e} s. - XX^{e} s. en décembre 2013, version qui s'inspirait des choix opérés dans le cadre de la GGHF[1], du moins pour les textes juridiquement ouverts. Cette version a été révisée en mai 2014 pour la période XVI^{e} s. - XVIII^{e} s.

Le travail accompli aurait été impossible sans l'aide efficace de M.-L. Demonet et L. Bertrand pour les Bibliothèques Virtuelles Humanistes (http://www.bvh.univ-tours.fr), de V. Montémont pour FRANTEXT (http://www.frantext.fr/), de R. Morrissey pour l'ARTFL (http://artfl-project.uchicago.educ).

Une convention de partenariat tripartite (ENS, Université de Cologne, ATILF pour FRANTEXT) est en passe d'être signée, qui encadrera la collaboration de PRESTO avec FRANTEXT.

Les critères qui ont guidé nos choix ont été les suivants :

- **Statut juridique ouvert** : L'objectif étant de pouvoir diffuser librement les textes du corpus noyau qui auront été étiquetés et lemmatisés par nos soins, tous les textes figurant dans le noyau i) soit possèdent un statut ouvert (cas des textes issus des BVH par ex.), ii) soit pourront évoluer vers un tel statut d'ici à la fin du projet. Un important travail (encore en cours) a été accompli avec nos collègues chercheurs et enseignants chercheurs des laboratoires impliqués dans les bases textuelles partenaires de PRESTO pour i) établir quels étaient les textes susceptibles de passer sous licence libre s'ils n'y figuraient pas, ii) pour identifier la licence Creative Commons qui pourrait permettre ce passage.
- **Graphies** : La qualité des éditions sélectionnées est un critère prépondérant. Les premières éditions des textes, notamment quand elles ont été publiées du vivant de l'auteur, ont été préférées ; à défaut, le choix s'est porté sur les éditions les moins interventionnistes, respectant l'orthographe d'époque. Lorsqu'une même édition s'est avérée disponible dans plusieurs bases textuelles, le choix a été fait de se tourner vers la base proposant le plus d'informations sur la fiabilité de la transcription et les choix éditoriaux opérés, et pour laquelle le statut juridique était le plus clair et le plus ouvert.
- **Répartition chronologique** : nous avons essayé, dans la mesure du possible, d'équilibrer le nombre des textes et/ou le nombre de mots pour tous les demi-siècles à partir du XVI^{e} s. Chaque fois que cela était possible (compte tenu des autres critères de sélections mis en jeu), nous avons poussé l'équilibrage jusqu'à

1Nous remercions S. Prévost (LaTTICe) qui nous a communiqué en 2013 la liste des textes sélectionnés dans le corpus de la GGHF, liste qui nous a été très précieuse pour la constitution de cette première version de notre corpus.

essayer de représenter toutes les décennies.

- **Equilibrage en genres/domaines** : bien que la réflexion sur les descripteurs genre et domaine ne soit pas encore achevée, nous avons essayé d'équilibrer sur les demi-siècles la représentation des différents genres textuels à partir des informations indiquées par les différentes bases textuelles consultées.
- **Equilibrage des formes** : nous avons essayé d'équilibrer la part respective des textes en prose et en vers, sachant que la prose prend une place croissante à partir du XII^{e} s.

A partir du corpus noyau ainsi constitué, et actuellement pour la seule période XVI^{e} s. – XVIII^{e} s. (inclus), un échantillon d'apprentissage a été constitué pour l'entraînement des outils.

1.3. Corpus second: constitution d'un corpus de presse diachronique

Une partie des travaux a été consacrée à la constitution d'un corpus de presse diachronique composé d'articles parus dans une dizaine de périodiques (nationaux et régionaux) au courant du 19e siècle (*Le Figaro*, *Journal des débats*, *La Presse* ; *L'Eclair*, *Le Passe-Temps*, *La Renaissance*, *Le Réveil de Lyon*) et au début des années 2000 (*Le Figaro*, *Le Monde*, *Libération* ; *L'Est républicain*, *Ouest-France*, *Sud Ouest*). A présent, les échantillons compilés existent sous forme de différents formats qui varient selon l'état des graphies issues des documents sources et le balisage d'unités structurelles. Le tableau suivant donne un aperçu des échantillons de périodique disponibles :

Échantillon (cote)	Périodique	Années	Graphies	Unités structurelles	Nombre de numéros disponibles
lfi19s-v01	Le Figaro	1826, 1827, 1830, 1835, 1840, 1855, 1857, 1860, 1865, 1870, 1875, 1880, 1885, 1890, 1895, 1896, 1898, 1901	OCR (BnF)[1] non corrigé	numéro, phrase	3 633
lfi19s-v02	Le Figaro	1826, 1827, 1830, 1835, 1840, 1855, 1860, 1865, 1870, 1875, 1880, 1885, 1890, 1895	OCR (BnF) corrigé	numéro, phrase	157
lfi19s-v03	Le Figaro	1826, 1830, 1835, 1840, 1855, 1860, 1865, 1870, 1875, 1880, 1885, 1890, 1895	OCR (BnF) corrigé	numéro, article, paragraphe, phrase	45
lfi02	Le Figaro	2002	archives numériques	numéro, article, paragraphe, phrase	313
jdd19s	Journal des débats	1830, 1840, 1850, 1868, 1870, 1883, 1884, 1892	OCR (BnF) non corrigé	numéro, phrase	2 234
la_presse19s	La Presse	1840, 1850, 1860, 1870, 1880, 1890, 1900	OCR (BnF) non corrigé	numéro, phrase	2 239
ecly19s	L'Éclair (Lyon)	1881, 1882, 1883, 1884, 1885	OCR (BML-COL)[2] non corrigé	numéro, phrase	220

pately19s	Le Passe-Temps (Lyon)	1879, 1889, 1890, 1891, 1892, 1893, 1894, 1895, 1896, 1897, 1898, 1899	OCR (BML-COL) non corrigé	numéro, phrase	575
renly19s	La Renaissance (Lyon)	1877, 1878, 1879, 1880, 1881, 1882, 1883, 1884	OCR (BML-COL) non corrigé	numéro, phrase	260
revly19s	Le Réveil de Lyon	1881, 1882	OCR (BML-COL) non corrigé	numéro, phrase	251
lm02	Le Monde	2002	archives numériques	numéro, article, paragraphe, phrase	312
souest02	Sud Ouest	2002	archives numériques	numéro, article, paragraphe, phrase	271
lestr02	L'Est républicain	2002	archives numériques	numéro, article, paragraphe, phrase	358

Après avoir été catégorisée et lemmatisée au moyen de *TreeTagger*, une partie des échantillons a été indexée et intégrée dans une base textuelle dont la composition est décrite par le tableau suivant :

Sous-corpus	**Mots-occurrences**
Journal des débats 1880-1900	12.152.000
La Presse 1880-1900	10.797.000
Le Figaro 1882-1901	33.714.000
L'Éclair (Lyon) 1881-1885	2.422.000
La Renaissance (Lyon) 1877-1884	2.163.000
Le Passe-Temps (Lyon) 1876-1899	4.130.000
Le Réveil de Lyon 1881-1882	3.369.000
Le Figaro 2002	26.995.000
Le Monde 2002	25.949.000
Est Républicain 2002 - version abrégée	26.004.000
Sud Ouest 2002	26.793.000

1.4. Opérations sur le corpus noyau en vue de sa lemmatisation dans PRESTO

1.4.1. Constitution d'un échantillon d'apprentissage

La lemmatisation et l'étiquetage morpho-syntaxique de textes antérieurs au XX^{e} s. est un des objectifs majeurs de PRESTO. Pour le réaliser au mieux, les informaticiens linguistes du projet ont noué des relations de collaboration avec d'autres chercheurs qui élaborent des ressources et des lexiques utiles au travail d'étiquetage et de lemmatisation. Ainsi, G. Souvay, ingénieur de Recherche CNRS à l'ATILF qui développe depuis plusieurs années LGeRM[2] (**L**emmes **G**raphies **e**t **R**ègles **M**orphologiques), lemmatiseur conçu pour gérer la variation graphique historique du français, participe-t-il étroitement à PRESTO, dans le cadre d'une convention tripartite ENS de Lyon, Univ. de Cologne, CNRS. De même M.-H. Lay, Maître de Conférences à l'Université de Poitiers et qui a conçu le logiciel ANALOG[3], est impliquée dans le processus d'annotation de l'échantillon et du corpus d'appprentissage (cf. s*upra*).

Dans une première phase du travail, nous avons décidé de mettre un focus particulier sur la tranche temporelle allant du XVI^{e} s. au XVIII^{e} s. Sur le plan linguistique, l'orthographe, le lexique et la syntaxe ayant évolué durant cette période, nous avons choisi d'y découper des sous-tranches temporelles en vue de construire des modèles linguistiques relativement stables. La consultation de collègues spécialistes de ces périodes[4] nous a conduit à adopter un découpage en 6 périodes :

1. **1500-1530** Au vu du nombre trop restreint de textes pour cette période celle-ci ne sera pas traitée dans cette première étape d'annotation.
2. **1530-1590/1610** Rem: l'édition de 1595 des *Essais* est considérée comme relevant de la période 2, et *l'Astrée* (édition de 1607 pour la 1ère partie) est considérée comme relevant de la période 3.
3. **1590/1610-1660**
4. **1660-1720**
5. **1720-1761**
6. **1761-1789**

Pour chacune de ces périodes sera construit un modèle linguistique créé à partir d'un échantillon d'apprentissage. Cet échantillon sera prélevé sur le corpus noyau à raison de 100 000 mots par période (la première période est écartée pour le moment), en équilibrant une fois de plus les genres et domaines, la forme et la répartition chronologique. Ces 500 000 mots seront désambiguisés et corrigés manuellement.

1.4.2. Calendrier des opérations en vue de la lemmatisation du noyau de PRESTO

- ***Étape 1*** **: Constitution d'un corpus d'apprentissage pour la période XVI^{e} s. - XVIII^{e} s.**
    - ***Étape 1-A*** : Construction d'un lexique morphologique de formes graphiques adaptées par archaïsation du lexique morphologique disponible pour le français dans le cadre de la boîte à outils FreeLing et en utilisant les règles morphologiques de LGeRM ; validation des formes produites.
    - ***Étape 1-B*** : Désambiguïsation par des experts linguistes des annotations obtenues par projection de la ressource dictionnairique sur les textes de l'échantillon d'appprentissage (500 000 mots (token)). Cette étape de vérification se fera au moyen de l'outil *Analog*

2 *LGeRM : lemmatisation de la variation graphique des états anciens du français et lexiques morphologiques* , ATILF - CNRS & Université de Lorraine. http://www.atilf.fr/dmf/LGeRM/. LGeRM (**L**emmes **G**raphies **e**t **R**ègles **M**orphologiques) a été initialement développé pour le moyen français (1330-1500) puis adapté au français du XVIe et XVIIe.

3 AnaLog permet d'interroger simultanément, en les rendant visuellement accessibles, des observables construits (corpus de textes – ordre syntagmatique) et des modèles de représentation (ressources d'annotation – ordre paradigmatique), afin d'étudier le résultat – toujours recomposable – issu du croisement des deux : les corpus annotés. (Lay, M.-H. & Pincemin, B. (2010)). Toutes les occurrences d'une erreur détectée sont localisées à l'aide d'un concordancier ; la modification (correction), effectuée sur le résultat de la concordance, porte en un seul temps sur toutes les occurrences similaires.

4 Ont été consultés M. Clément (Univ. Lyon 2), M.-L. Demonet (Univ. Rabelais, Tours), N. Fournier (Univ. Lyon 2), S. Rémi-Giraud (Univ. Lyon 2), D. Reynaud (Univ. Lyon 2).

développé par M.-H. Lay (Univ. Poitiers http://ll.univ-poitiers.fr/sdl/spip.php?article15). Une première campagne d'annotation est programmée durant la dernière semaine de juin 2014.

- ***Étape 2 :*** Création de modèles d'étiquetage automatique et évaluation
  - **Étape 2-A** : Création de modèles linguistiques par apprentissage sur la base des textes contenant les annotations désambiguïsées par les experts. Il est prévu de créer des modèles pour différents étiqueteurs (TreeTagger, MElt, Morfette, etc).
  - **Étape 2-B** : Évaluation des modèles après leur application sur un corpus de test extrait du corpus d'apprentissage. Identification et analyse des erreurs en relation avec les experts.
  - **Étape 2-C** : Avis définitif sur la fiabilité des modèles. En cas d'une qualité jugée suffisante, choix du meilleur modèle, sinon réajustement des paramètres d'apprentissage et reprise des étapes 2-A et 2-B.
- ***Étape 3*** : Lemmatisation du corpus noyau de PRESTO, vérification et correction par des experts.
  - ***Étape 3-A*** : Application du meilleur modèle d'étiquetage et de lemmatisation à l'ensemble du corpus noyau de PRESTO
  - ***Étape 3-B*** : Vérification par des experts linguistes et correction des sorties obtenues au moyen des outils automatiques d'étiquetage et de lemmatisation appliqués au corpus noyau. Version définitive de ce corpus.
  - ***Étape 3-C*** : Les textes avec leurs annotations vérifiées et corrigées par les experts peuvent être réinjectés dans les outils automatiques afin d'améliorer leurs performances.

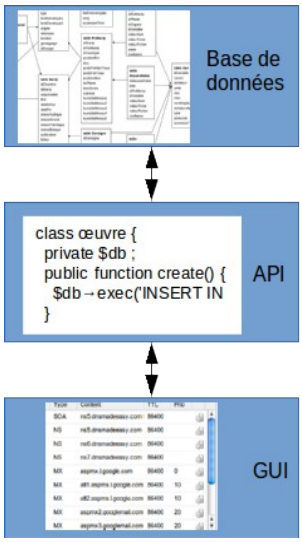
La première campagne d'annotation correspondant à l'étape 1-B du calendrier ci-dessous a été lancée durant la dernière semaine du mois de juin. Le jeu d'étiquette PRESTO14 a été mis au point et le travail d'annotation proprement dit a commencé. Nous espérons que l'annotation automatique de l'échantillon d'apprentissage sera entièrement vérifiée manuellement en décembre 2014.

2. Base de données, interfaces et outils d'exploration.

2.1 Architecture

Créer une base à la fois simple et puissante représente un défi, que l'on peut aisément s'épargner. Ainsi, pour le projet Presto, on distingue la *base de données* proprement dite, qui est régie par des considérations informatiques de « puissance » (expressivité et performances), et les *interfaces* d'utilisation de cette base, qui doivent quant à elles être simples à utiliser.

Figure 1: Architecture de la base Presto et de ses interfaces



On distingue deux interfaces : une interface informatique (API), et une interface graphique (GUI). Ces deux types d'interfaces exposent des fonctionnalités d'édition classique (CRUD[5] : *Create*, *Read*, *Update*, *Delete*).

Ces composants sont organisés selon une architecture classique à trois niveaux[6], qui permet une certaine souplesse, car chacun de ces composants sera amené à évoluer au fil du projet. La GUI fait appel à l'API qui elle même fait appel à la base de données (ainsi la GUI ne communique jamais directement avec la base).

2.2. Structure de la base de données

La base est structurée de manière hiérarchique. À une *œuvre*, correspondent une ou plusieurs *éditions*, qui sont constituées d'une succession ordonnée de *parties* (avertissement, préface, texte proprement dit, qui peut être découpé en plusieurs tomes ou parties, etc.), constituées d'une suite d'*échantillons*,

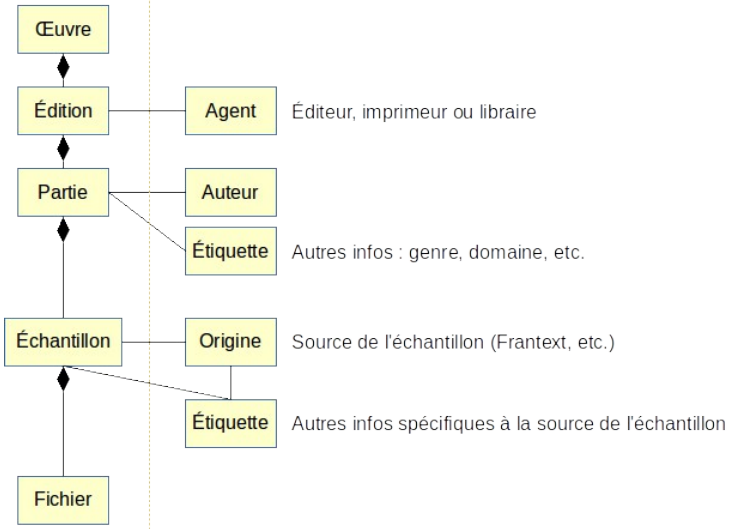
5 http://fr.wikipedia.org/wiki/CRUD
6 http://fr.wikipedia.org/wiki/Architecture_trois_tiers

finalement constitués de *fichiers*.

Chaque niveau se voit associer des *étiquettes*. Par souci d'expressivité (au détriment de la clarté), celles-ci sont définies au niveau le plus « bas ».

Par exemple, l'*auteur* est défini dans la base au niveau de la *partie*, et non au niveau de l'*œuvre* car il peut y avoir des auteurs différents pour chaque partie (paratexte, œuvres collectives, etc.). La notion d'auteur d'une œuvre n'existe donc pas explicitement dans la base. Toutefois, l'utilisateur s'attend à avoir un auteur pour une œuvre. C'est pourquoi les interfaces introduisent un auteur « virtuel » (ce caractère virtuel n'étant pas connu de l'utilisateur), défini comme la liste des auteurs des parties non paratextuelles d'une œuvre.

Figure 2: Structure (simplifiée) de la base de données.



Avancement au 20 mai 2014

- Base : 80 % (manque les niveaux échantillon et fichier)
- API CRUD: 40 %
  - Create : 80 % (manque les niveaux échantillon et fichier)
  - Read : 80 % (manque les niveaux échantillon et fichier)
  - Update : à faire
  - Delete : à faire
- GUI[7] : 25 %
  - Create : 20 % (importation de fichiers CSV, pas encore de création directe)
  - Read : 80 % (manque les niveaux échantillon et fichier)
  - Update : à faire
  - Delete : à faire

Remarques

À noter, le type *Date* de MySQL ne convient pas pour le projet, car il ne supporte que les dates sur l'intervalle 1er janvier 1000 – 31 décembre 9999. Les années sont donc stockées sous forme d'entiers signés sur 2 octets (intervalle : -32768 à 32767).

2.3. Lemmatisation

Par « lemmatisation », on entend en fait un processus complet de :

1. tokenisation, d'après une liste de séparateurs prédéterminés ;
2. annotation en lemmes ;
3. annotation morpho-syntaxique (parties du discours).

L'annotation se sert du lexique LGeRM adapté au projet, et du « devineur » (*guesser*) de LGeRM pour les

7 http://presto.aiakide.net/bd/ (adresse temporaire, login/mdp requis : guest/pr3sto pour accès en lecture seule)

mots inconnus, accessible via un service Web REST.

Aucune désambiguïsation n'est effectuée : toutes les annotations possibles sont indiquées.

Certains textes comportent déjà une lemmatisation (balise *w*) ; afin d'évitet tout conflit, on utilise donc un autre nom de balise (*wpresto*). Exemple de sortie du lemmatiseur :

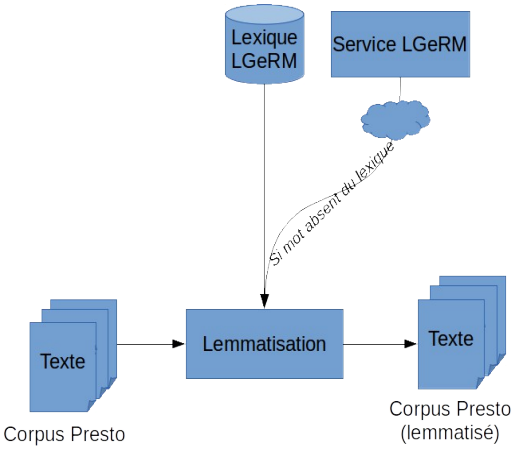


Figure 3: Ressources utilisées pour la lemmatisation.

```
<wpresto dico="agamemnon:NP">*Agamemnon</wpresto><lb/>
<wpresto dico="OH:I">Ô</wpresto>
<wpresto dico="VIEILLARD:NCMS000">vieillard</wpresto>,
<wpresto dico="HÂTER:VMIP1P0|HÂTER:VMM01P0">hâtons</wpresto>
-
<wpresto dico="NOUS:PP1CP000">nous</wpresto>
:
<wpresto dico="LE:DA0CS0|LE:PP3CSA00">l</wpresto>
'
<wpresto dico="HEURE:NCFS000|HURE:NCFS000">heure</wpresto>
<wpresto dico="FUIR:VMIP3S0|FUIR:VMIS3S0|FUIR:VMSI3S0">fuit</wpresto>
.
```

Avancement au 20 mai 2014

- Fait :
  - identification des noms propres Frantext (commencent par une *)
  - identification des nombres arabes
- En cours :
  - identification des mots coupés (césure)
- À faire :
  - identification des lexèmes discontinus
  - identification des numériques romains

Perspectives

Après désambiguïsation automatique, les hypothèses de lemmatisation seront triées par probabilité décroissante, et on indiquera le score de désambiguïsation de chaque hypothèse.

2.4. Plateforme BTLC et boîte à outils *PrimeStat*

La plateforme web BTLC donne accès à différentes applications permettant d'explorer les bases textuelles constituées dans le cadre du projet PRESTO. Ces applications sont implémentées par la boîte à outils *PrimeStat* qui comprend les fonctionnalités suivantes :

- création de sous-corpus ou de partitions selon les méta-données liées aux textes faisant partie des bases disponibles
- définition d'expressions de requête complexes en termes de propriétés lexicales ainsi que du nombre de segments ciblés
- extraction et manipulation de concordances KWIC
- création d'index et calcul de spécificités fréquentielles
- calcul de cooccurrences lexico-syntaxiques
- analyses multivariées
- AFC
- Classification ascendante hiérarchique
- Classification K-Means

2.5. Plate-forme TXM et boîte à outils

On se reportera à l'URL: http://textometrie.ens-lyon.fr/spip.php?rubrique96

3. Publications et événements

Outre la participation de membres de PRESTO à des événements (journées d'étude, conférences, … : la liste est accessible sur le site http://presto.ens-lyon.fr), les publications déjà acceptées ou soumises sont les suivantes:

Blumenthal, P. (à par. 2014), « Caractéristiques et effets de la complexité sémantique de noms d'affect » in Nouvelles perspectives en sémantique lexicale et en organisation du discours, I. Novakova & P. Blumenthal (eds).

Blumenthal, P. (soumis), « L'expression prépositionnelle de la simultanéité au 20e siècle ».

Royer, L., Vigier, D. (à par. 2014b), « Les collocatifs nominaux des prépositions *en, dans, dedans* au XVIe s », in *Nouvelles perspectives en sémantique lexicale et en organisation du discours*, I. Novakova & P. Blumenthal (eds).

Vigier, D. (à par. 2015), « Les prépositions *en, dans* et *dedans* au XVIe s. Approche statistique et combinatoire», *Le Français Moderne*.