

Pour une nouvelle philologie numérique : réflexions sur la relation texte(s)-document(s)

St. Polis

(F.R.S.-FNRS – Université de Liège)

B. Stasse

(Université de Liège)

Résumé. Deux axes ont été privilégiés dans le développement des bases de données textuelles qui, depuis l'avènement de la micro-informatique, ont été constituées afin de faciliter et d'enrichir l'étude des documents en langues anciennes. Le premier correspond à l'encodage, dans une version standardisée, de textes reproduisant une édition particulière ; ceux-ci sont aisément interrogeables, mais la dimension documentaire est systématiquement négligée : les informations relatives à la tradition manuscrite et, plus largement, au processus éditorial, sont absentes. Le second axe, quant à lui, intègre exclusivement la dimension du document. Il demeure quantitativement (à la fois du point de vue du nombre de projets et de la quantité de textes concernés) peu représenté et qualitativement (type de données encodées, moteurs de recherche disponibles, etc.) non satisfaisant. Cette contribution se propose, par la combinaison de ces deux axes, d'envisager les « textes » et les « documents » dans leur complémentarité. En proposant une modélisation de la relation intrinsèquement simple qu'ils entretiennent, elle se veut une première étape du nécessaire processus d'interrogation de nos pratiques en matière de philologie numérique.

Introduction

Les textes sont, à n'en pas douter, l'une des sources privilégiées pour notre connaissance des civilisations du passé. Cela justifie assez que toutes les disciplines aient cherché à y trouver matière pour leurs études et qu'elles aient développé des outils propres en vue de les appréhender avec une méthodologie

satisfaisante. Ces textes, comme objets empiriques, délimitent par conséquent un domaine de recherche naturellement interdisciplinaire¹ ; plus largement, Fr. Rastier (2001 : 1-2) souligne le fait qu'« [i]ls intéressent passablement les linguistes et les psycholinguistes, beaucoup les “littéraires”, et bien d'autres, en didactique, en informatique, bref les gens du texte, épars dans diverses diasporas et divisés en maintes tribus académiques. » Et l'auteur d'ajouter que, à son sens, « une synthèse semblerait prématurée sinon illusoire, car l'ensemble des textes compose un domaine d'objectivité dans lequel chaque discipline élit domicile à sa guise. »

Il est certain que la richesse réflexive qu'a pu engendrer la rencontre de différentes disciplines au sein des textes n'a eu d'égal que le flou dans lequel a été laissée la notion elle-même² ; il faut de même reconnaître que les chercheurs qui l'utilisent, bien que travaillant sur un objet commun, se distinguent par leurs questionnements, par les objectifs qu'ils poursuivent, par les méthodes appliquées et souvent par le statut épistémologique de leur discipline. Toutefois, il semble que la majorité des problèmes posés par la définition du texte et par l'objectivation de cette dernière peut être levée si l'on prend soin de penser le texte en complémentarité avec sa face matérielle, le document.

Parmi les sciences du texte, la philologie et la linguistique illustrent clairement les deux perspectives complémentaires qui peuvent être mobilisées autour du texte — le fait que la seconde soit, en quelque sorte, une émanation de la première, n'est certainement pas étranger à ce phénomène. Elles sont ainsi propres à jeter quelques lumières sur la notion de texte et ont servi directement de base aux réflexions qui suivent.

Historiquement, la philologie (dont on peut situer les origines à l'époque de la première sophistique), a toujours été indissociable de la grammaire : depuis l'école d'Aristarque liée à la bibliothèque d'Alexandrie et la *Technè* de Denys le Thrace, jusqu'aux éminents philologues de l'école allemande du XIX^e siècle en passant par les A. Politien, L. Valla et autre Pic de la Mirandole, nul n'a songé à remettre ce lien en question. Au cours du XIX^e siècle, l'émergence de la linguistique comme discipline autonome ambitionnant d'étudier la langue

1 On verra en ce sens les différents points de vue retenus pour répondre à la question « qu'est-ce qu'un texte ? » dans E. Barbotin [dir.] (1975).

2 Voir les définitions analysées par Fr. Rastier (1996 : 16-21).

pour elle-même est allée de pair avec une rupture (que prépare à sa manière le *Cours de linguistique générale*³) d'avec la philologie : progressivement, l'étude des langues n'a plus nécessairement été envisagée de concert avec l'étude des textes⁴.

Les nouvelles technologies de l'information permettent aujourd'hui de dépasser ce clivage et paraissent autoriser un traitement simultané des niveaux philologique et linguistique. Pour ce faire, il convient de proposer un modèle⁵ qui rende justice à la complexité de la relation entre textes et documents au niveau philologique et permette, dans le même temps, des analyses linguistiques

3 La rupture entre philologie et linguistique n'était pas un objectif pour F. de Saussure (voir *e.g.* 1972 [1916] : 351), mais est probablement la conséquence du problème non résolu du lien unissant langue et langue discursive (ou parole). Il constatait par ailleurs que « [l]a langue n'est pas l'unique objet de la philologie, qui veut avant tout fixer, interpréter, commenter des textes. La philologie use de sa méthode propre qui est la critique. Si elle aborde les questions linguistiques, c'est surtout pour comparer les textes de différentes époques, déterminer la langue particulière à chaque auteur, déchiffrer et expliquer des inscriptions rédigées dans une langue archaïque obscure » (F. de Saussure 1972 [1916] : 13-14).

4 L. Hjelmslev conserve le rare mérite (cf. Fr. Rastier 2001 : 61) de considérer que la théorie du langage est une théorie du texte (« [i]l est certain que l'analyse du texte [...] échoit au linguiste comme une obligation irréductible », L. Hjelmslev 1968-1971 : 28) ; pour les difficultés posées par le concept de « texte » chez L. Hjelmslev, voir S. Badir (1998) et R. Kyheng (2005). Sa position diffère donc de celle des linguistes ayant théorisé la phrase comme une unité de discours, en dehors de la linguistique *stricto sensu*. On connaît les positions adoptées par des figures éminentes comme G. Guillaume ou Ém. Benveniste en la matière. En des termes semblables, ils ont repoussé la phrase au-delà du dernier niveau de l'échelle des combinaisons linguistiques codées ; on quitte ainsi, avec la phrase, la langue comme système de signes. Méthodologiquement, cette perspective s'explique, mais elle n'a rien d'évident : « [s]i les textes sont l'objet empirique de la linguistique, isoler des phrases, *a fortiori* des mots, résulte de procédures méthodologiques plus complexes qu'il ne semble » (Fr. Rastier 2001 : 31). Sur les enjeux de la linguistique textuelle, on peut se reporter commodément à O. Soutet (2001 : 323-346) et J.-M. Adam (2004 [1999]).

5 À propos des sciences du texte, J.-M. Viprey (2006) note que « si le numérique est bien le vecteur, le siège et l'enjeu d'une révolution technique, il serait aussi absurde qu'il l'a toujours été, et combien plus dangereux que jamais, de croire que cette révolution technique irait engendrer spontanément une mise à jour théorique, ainsi que de s'illusionner sur la qualité des réponses techniques primitives, surgies du marché des "nouvelles technologies" ».

multiplés (tant dans leur variété que dans leur portée)⁶ ; un enrichissement mutuel en découlerait nécessairement : la philologie pourrait bénéficier des progrès de l'étude des langues dans son travail critique et la linguistique retrouver l'assise empirique dont elle a besoin⁷. En outre, ce modèle devrait idéalement permettre de préciser, sinon de raffiner, les approches du texte relevant d'autres niveaux d'analyse (rhétorique, stylistique, etc.).

Force est toutefois de constater que malgré des progrès notables dans le domaine informatique et l'essor considérable de la linguistique de corpus (qui se divise aujourd'hui en plusieurs spécialités en fonction des objectifs poursuivis), la philologie numérique ne possède pas — quoi qu'en pensent certains — de véritable assise méthodologique. Parallèlement, l'analyse linguistique ne fait guère de place à la dimension textuelle des corpus étudiés. Il est pourtant indéniable qu'une prise en compte de la variante (dans toutes les acceptions conceptuelles du terme) permettrait de rendre aux variations de la parole l'importance qui leur est due.

Il faut constater que les bases de données et corpus annotés sur lesquels repose désormais largement l'étude des textes en langues anciennes — que l'on se situe dans le domaine classique ou oriental — sont caractérisées par une absence frappante d'intégration de la relation qu'entretiennent « texte(s) » et « document(s) ». Si elles permettent souvent une exploration fine du texte, la rupture d'avec le document semble généralisée, lors même que les disciplines desquelles émanent ces bases de données constituent le berceau de la philologie et, partant, envisagent la pratique du document avec rigueur autant que respect. Ce constat, aussi abrupt qu'indéniable, impose de s'interroger sur les motifs qui ont présidé à la création d'outils ignorant la dimension documentaire d'un texte (sinon textuelle d'un document). Si la dimension technique — entendue ici au sens large de possibilité d'implémentation — peut être invoquée pour les initiatives les plus anciennes, il apparaît que l'inadéquation

6 Tout récemment (cours 2008-2009), Fr. Rastier a plaidé en faveur d'un « agenda épistémologique et méthodologique » dont le premier point est « un moratoire sur les modèles partiels, pour restituer la complexité des textes » (cf. <http://www.ehess.fr/fr/enseignement/enseignements/2008/ue/3683/>).

7 En poursuivant un objectif de scientificité parfois mal compris, certaines écoles linguistiques ont cherché à s'écarter des humanités ; elles ont ainsi renoncé à la perspective critique qu'elles auraient pu hériter de la philologie.

détermination de cette relation et des termes qu'elle implique n'est assurément pas étrangère à cet état de fait.

Après avoir esquissé un panorama critique de différents projets existants qui ont pour objectif l'encodage structuré de ressources textuelles (§ 1), on s'attachera à fournir une série de propositions définitives (§ 2) afin d'objectiver les termes impliqués dans cette relation et, dès lors, d'en permettre une modélisation qui satisfasse à nos pratiques du document (§ 3).

1. *Status quaestionis*

Comme on l'a dit, la philologie numérique nécessite comme préalable la modélisation de la relation texte(s)-document(s). Dans la pratique, elle peut reposer sur des normes internationales d'encodage des textes⁸ et sur des travaux bien établis dans les domaines de la constitution et de l'analyse de corpus⁹. Ces derniers travaux ont été réalisés selon deux orientations principales. La première se concentre sur le texte (§ 1.1) et vise à élaborer des corpus en formant un ensemble homogène — temporel, géographique, thématique, etc. —, ou en recherchant la quantité, voire l'exhaustivité. S'ils ouvrent la voie à tout un champ d'études sur les textes, ils sont généralement réalisés au prix d'une normalisation de l'information. La seconde orientation se concentre sur le témoin (§ 1.2) et vise, essentiellement, à faciliter son encodage et sa diffusion¹⁰.

8 Une grande confusion a longtemps régné en matière d'encodage de données textuelles, restreignant d'autant leur portabilité. Diverses tentatives de normalisation ont ainsi vu le jour. Apparu au milieu des années 60 aux États-Unis, l'ASCII a longtemps été la seule norme de codage universellement utilisée. Codé originellement sur 7 bits, l'ASCII permet le codage de 128 caractères. À la suite, de nombreuses normes se sont développées sur une structure 8 bits, portant la limite à 256 caractères (ainsi la série ISO-8859, déclinée en 16 normes, dont l'ISO-8859-1 pour l'Europe occidentale, et l'ISO-8859-7 pour le grec). Afin de remédier aux problèmes liés à cette multitude de normes et de standards (recours à plusieurs normes pour une même langue, incompatibilité entre les différents systèmes de codages), le standard Unicode a été développé, qui code les caractères sur un maximum de 32 bits et attribue un numéro unique à chaque caractère, abstraction faite des plateformes, langues ou logiciels utilisés. Sur le standard Unicode, voir P. Andries (2008) ; <http://unicode.org/>.

9 Voir, par exemple, B. Habert, C. Fabre & F. Issac (1998) ainsi que les différentes contributions à la revue *Corpus* (<http://corpus.revues.org>).

10 Sur les définitions retenues dans ce travail pour les notions de « document », « texte », « témoin » et « unité textuelle », cf. *infra* § 2.

1.1. Le texte comme objet

Les projets qui envisagent le texte comme objet sont le fruit d'une longue tradition qui s'enracine dans les sciences de l'Antiquité. Parmi les chercheurs en sciences humaines, les antiquisants ont en effet été les premiers à tirer parti des possibilités nouvelles qu'offrait à leur discipline l'outil informatique¹¹.

Ainsi, la tradition considère comme le fondateur de la recherche linguistique par ordinateur le Père Busa qui, dès 1949, a utilisé les calculateurs IBM pour réaliser des relevés lexicographiques dans l'œuvre de saint Thomas d'Aquin¹². Son premier index, réalisé à partir de cartes perforées, paraît en 1951. Les premiers traitements de la langue grecque apparaissent dès 1956 avec L. Brandwood qui, s'attachant particulièrement aux relevés fréquentiels, s'inscrit dans le courant de la statistique littéraire. En 1961 est créé, premier en son genre, le Laboratoire d'analyse statistique des langues anciennes à l'Université de Liège¹³.

L'idée suivie dans cette orientation est de recourir à la fonction première de l'ordinateur, à savoir ordonner des données nombreuses pour en faciliter le traitement. Les premiers corpus traités étaient constitués de sources littéraires ; dans le domaine des sciences de l'Antiquité, les épigraphistes et papyrologues étaient alors tenus éloignés par les restrictions techniques des premiers temps (impossibilité de saisir des graphèmes complexes et des signes critiques nombreux ; problème de l'impression des résultats)¹⁴.

À seule fin de proposer des instruments pour étudier les textes, les travaux se distinguent par la perspective adoptée qui régit la nature du texte encodé.

1.1.1. Le texte brut

Une première méthode consiste à rassembler dans un format numérique un corpus de textes le plus exhaustif possible, l'interrogation d'un très large

11 Pour un aperçu de l'histoire de l'informatique appliquée aux langues anciennes, voir J. Denooz (2007).

12 Travaillant sur le vocabulaire de l'intériorité chez saint Thomas d'Aquin, Busa jugea insuffisants pour son étude les index et concordances existants, qui excluaient les mots très fréquents (prépositions, coordinations, etc.) ; voir Th.N. Winter (1999 : 3-21).

13 Voir <http://www.cipl.ulg.ac.be/bdla/la/>.

14 C'est pour ces mêmes raisons que la seconde orientation, tournée vers le témoin, ne fut envisagée que plus tard.

corpus autorisant une série d'analyses nouvelles (études linguistiques ; nouvelles éditions ; etc.). Les requêtes s'effectuent sur un texte figé qui est la reproduction du texte d'une édition de référence¹⁵ et les dimensions « document » et « témoin » sont absentes de ces corpus¹⁶.

Les principaux représentants de cette catégorie sont, dans le domaine des langues classiques, le *Thesaurus Linguae Graecae* (TLG) pour la littérature grecque, et les deux bases du *Packard Humanities Institute* (PHI) pour la littérature latine d'une part, et l'épigraphie et la papyrologie grecques d'autre part¹⁷.

Lancé en 1972 par Th.F. Brunner (University of California, Irvine), le TLG propose aujourd'hui la quasi totalité des textes grecs écrits entre le VIII^e et le VI^e, et la plus grande partie de ceux écrits jusqu'en 1453. L'encodage se poursuit actuellement, le but étant de fournir une banque de données de la littérature grecque de l'Antiquité à nos jours. Le corpus fait actuellement état de 99 millions de mots, issus de 12 000 œuvres rédigées par près de 3800 auteurs¹⁸. L'ensemble se présente sous la forme de fichiers textes codés en Betacode¹⁹.

Réalisée à partir de 1987 par le *Packard Humanities Institute* (Californie), la banque de données du PHI 5 couvre l'essentiel de la littérature latine classique.

15 Il est à noter que le choix de l'édition peut se faire sur la base de critères n'étant pas exclusivement scientifiques, comme dans le cas d'un accord avec une maison d'édition.

16 On peut également citer, en marge de cette catégorie, des cas plus particuliers, comme la version électronique complète d'une édition critique spécifique. On en donnera comme exemple la version électronique des *Fragmente der griechischen Historiker* de F. Jacoby : il s'agit de la numérisation d'une édition donnée, à ceci près que, d'une part, cette édition forme un corpus cohérent en soi (l'ensemble des fragments des historiens grecs) et que, d'autre part, la numérisation reproduit l'ensemble des données de l'édition imprimée (avec introduction, notes, commentaires et appareil critique). La dimension restreinte du corpus et les possibilités de recherches basiques n'autorisent cependant pas d'étude linguistique approfondie et limitent l'intérêt de cette formule à la relative commodité de la consultation électronique d'un ouvrage.

17 D'autres initiatives similaires existent : entre autres, la *Bibliotheca Teubneriana Latina* (CD-Rom), réalisée par le CTLO, propose le texte, dans l'édition de la collection Teubner, de plus de 900 œuvres dues à 450 auteurs, couvrant la période 300^a–200^p (auxquels sont ajoutés quelques auteurs tardifs jusqu'en 500^p) ; la *Clauss Datenbank*, base de données en ligne des inscriptions latines, rassemble à ce jour plus de 460 000 inscriptions, issues de plus de 800 publications (voir <http://www.manfredclaus.de/>).

18 Le site du TLG est accessible à l'adresse <http://www.tlg.uci.edu/>.

19 Sur le Betacode, créé pour permettre le codage du grec polytonique au moyen des seuls caractères ASCII, voir <http://www.tlg.uci.edu/BetaCode.html>.

Ce même Institut a également créé le *PHI* 7, une banque de données rassemblant quelques 180 000 inscriptions grecques et 60 000 papyri grecs (représentant près de 60% de la documentation connue aujourd'hui).

On notera que, si les textes n'intègrent pas d'étiquette dans leur encodage original, des logiciels élaborés pour interroger ces banques de données créent aujourd'hui des liens vers (ou intègrent) des outils externes d'aide à la lecture des textes (dictionnaires, outils d'analyse morphologique²⁰). Étant donnée l'absence de marquage des textes, ces liens fonctionnent — cela va sans dire — à sens unique, de la forme du texte vers un ou plusieurs lemmes du dictionnaire, sans qu'il puisse s'agir d'analyse de formes en contexte²¹.

Ces exemples montrent suffisamment que, dans le domaine des langues classiques — et cela est vrai dans d'autres —, les banques de données textuelles sont aujourd'hui largement constituées. Aussi la question est-elle moins de savoir si un texte est encodé, que celle de savoir exactement quelles données ont été encodées.

1.1.2. Le texte avec étiquetage : les étiquettes explicatives

Si la finalité dans cette seconde catégorie est similaire — proposer des instruments pour l'étude de textes —, la perspective adoptée est différente : il ne s'est plus agi, dans ces projets, de fournir un corpus aussi étendu que possible, mais un corpus plus approfondi — et, pour cette raison, dans un premier temps au moins, plus restreint.

Le Laboratoire d'analyse statistique des langues anciennes (LASLA) de l'Université de Liège a joué dans ce domaine de l'informatique appliquée aux données textuelles un rôle de pionnier²². Le résultat est aujourd'hui une base

20 Voir par exemple les outils proposés par la *Perseus Digital Library* : <http://www.perseus.tufts.edu/hopper/>.

21 Même sans étiquetage d'origine, la possibilité pourrait néanmoins exister d'affiner l'analyse morphologique en l'adaptant au contexte par l'intégration d'un algorithme de désambiguïsation des formes.

22 Cf. *supra*. Le LASLA est à l'origine de la création des centres informatiques des autres institutions de la Communauté française, le Centre de traitement électronique des documents de Louvain (CETEDOC, depuis 1968) et le Groupe d'informatique et de traitement informatique de l'ULB (GITA, depuis 1974).

de données²³ qui propose au sein d'un corpus comptant près de deux millions de mots, les textes latins de grands auteurs classiques, avec, pour chaque terme, une analyse lexicologique, morphologique (pour chaque forme, une étiquette rassemble les indications du lemme, de la référence et de l'analyse morphologique complète) et en outre, pour les verbes, certains éléments d'analyse syntaxique (touchant à la syntaxe des propositions)²⁴. L'outil offre la possibilité de requêtes complexes (par lemme, subordination et catégorie grammaticale).

L'ensemble de textes ainsi constitué permet de réaliser des analyses linguistiques fines, notamment avec l'aide du logiciel Hyperbase, qui ajoute aux possibilités de requêtes des fonctions complexes d'exploitations statistiques²⁵.

Dans le domaine de la langue égyptienne, on mentionnera le *Thesaurus Linguae Aegyptiae* (TLA) de l'Académie des Sciences de Berlin-Brandenburg²⁶. Ce projet combine une base de données lexicale — elle repose sur les dépouillements réalisés pour la version imprimée du *Wörterbuch der Aegyptischen Sprache*²⁷ — avec un corpus textuel (*Textcorpus*) qui est progressivement enrichi de nouveaux textes. Chaque mot-forme des textes encodés est lemmatisé (c'est-à-dire qu'il est directement lié à une entrée du *Wörterbuch*), mais aucune analyse morphologique n'est proposée²⁸.

Tous les projets inscrits dans cette première orientation (qui se concentre sur l'objet texte), ont donc donné naissance à de précieux outils — devenus indispensables dans nos études —, mais conceptuellement lacunaires, dans cette mesure où la dimension du document demeure absente.

23 La base de donnée du LASLA est accessible à l'adresse <http://cipl.philo.ulg.ac.be/app/lasla/chkpass.html>.

24 Sur la méthode de traitement des formes, voir J. Denooz (1978 : 1-36).

25 Sur le logiciel Hyperbase appliqué à la banque de données du LASLA, voir S. Mellet & Ét. Brunet (<http://www.unice.fr/bcl/index2.php>).

26 Cf. <http://aew.bbaw.de/tla/index.htm>.

27 Le TLA donne ainsi accès, sous la forme de fiches scannées, aux *Zetteln* originaux qui ont servi de base au *Wörterbuch* de A. Erman & H. Grapow [éd.] (1926-1961).

28 Le Projet Ramsès de l'Université de Liège (voir S. Rosmorduc, St. Polis & J. Winand [2008] ; J. Winand, St. Polis & S. Rosmorduc [2009 – sous presse]) entend, pour le néo-égyptien, pallier cette lacune en intégrant l'ensemble des niveaux d'analyse linguistique (du morphologique au sémantique). De plus, pour le niveau graphique, il inclut les hiéroglyphes, lors que les textes du TLA sont simplement translittérés.

1.2. Le témoin comme objet : les étiquettes descriptives

Un autre point de départ a été de considérer les problèmes posés par les contraintes du numérique au rendu intégral d'un témoin. Si l'encodage d'un texte brut (ou presque²⁹) était suffisant dans le cadre des projets que nous venons d'évoquer, d'autres se sont interrogés sur la manière d'encoder l'édition diplomatique d'un texte ou, plus simplement, de diffuser commodément des textes pourvus de signes critiques. L'encodage des textes s'accompagne dans ce cas d'étiquettes descriptives.

La *Text Encoding Initiative*³⁰, projet lancé par un consortium en 1987, s'inscrit dans ce cadre : son objectif principal est le développement d'un standard pour la représentation fidèle des textes dans un format numérique — en d'autres termes, un système commun de codage de textes —, comprenant à la fois la description physique du document et la structure logique du texte³¹. L'idée première était de permettre la retranscription intégrale de tout témoin dans un format numérique (« *encoding of all kinds of features of all kinds of texts* »³²), et l'échange commode entre utilisateurs (« *standard format for data interchange* »³³).

Dépassant l'objectif premier d'affichage et de partage, les possibilités de balisage ont été étendues pour permettre l'enregistrement d'informations complémentaires, non plus nécessairement liées à la présentation particulière d'un document. Parallèlement, le recours à un encodage standardisé a ouvert la voie à la création de nombreux outils qui permettent de traiter, dans de nouvelles applications, les textes ainsi encodés (moteurs de recherches aux fonctions avancées, outils d'analyse statistiques, etc.).

Les projets relevant de cette seconde orientation, qui se concentre sur l'objet témoin, ont, par leur approche différente, jeté les bases nécessaires à un

29 Par exemple, on retrouve dans le *PHI* 7 quelques indications formelles, comme la présence de signes critiques de restitution et l'indication des lettres de lecture douteuse. Il en va de même pour les textes du *TLA*.

30 Voir <http://www.tei-c.org>.

31 « *Guidelines have been written largely with a focus on text capture (i.e. the representation in electronic form of an already existing copy text in another format) rather than text creation (where no such copy text exists)* » (L. Burnard & S. Bauman [éd.] 2008 : xxv–xxvi).

32 L. Burnard & S. Bauman [éd.] (2008 : xxv).

33 *Ibid.*

encodage numérique globalement cohérent. Cela étant, la conception reste lacunaire, qui n'intègre pas pleinement la dimension texte dans sa composante linguistique.

2. Propositions définitoires : le document, le texte et le témoin

Afin d'envisager la complémentarité entre texte(s) et document(s), il nous paraît que trois notions primaires doivent être mobilisées : le document, le texte et le témoin (cf. définitions 1 à 3). Chacune d'entre elles sera définie en relation avec les deux autres ; il s'agit donc d'interdéfinir les éléments d'une relation complexe³⁴. Pour ce faire, on suivra un parcours définitoire³⁵ qui passera successivement du concret à l'abstrait : en première approximation, on peut considérer que le document est le support physique d'un texte, sorte d'abstraction qui s'actualise sous la forme d'un ou plusieurs témoins. Ces trois notions sont utilisées dans l'élaboration conceptuelle d'un « dossier » qui assure l'unité et l'intégrité de la relation autour de la notion d'unité textuelle (cf. définition 4). De notre point de vue, l'aboutissement de la philologie numérique est ce dossier, qui sera l'entité encodée.

Définition 1. On nommera “document” « tout support matériel³⁶ sur lequel un (ou plusieurs) témoin(s) est (sont) conservé(s). »

Cette première définition n'appelle guère de commentaires tant elle recoupe le sens commun du terme « document »³⁷. Elle n'en possède pas moins certaines implications qu'il importe de souligner. (1) Comme « support matériel », le document est un médium qui possède, au minimum, les caractéristiques

34 Méthodologiquement, ce procédé ne confine pas à la circularité ; il suffit pour s'en convaincre de songer à la définition des relations sémiotiques.

35 Les propositions définitoires formulées voudraient respecter, autant que faire se peut, les définitions intuitives des notions envisagées, mais être suffisamment rigoureuses pour déboucher sur une modélisation stricte.

36 Sur l'importance de dépasser l'opposition « écrit » vs « oral » par la notion de support, cf. Fr. Rastier (2001 : 21).

37 Le signifié du terme est évidemment plus large, mais pour notre propos, il a été volontairement restreint aux supports matériels conservant un (ou plusieurs) témoin(s) de texte(s) ; tous les autres types de documents sortent du cadre ici considéré.

suivantes : un ou plusieurs nom(s)³⁸ ; une provenance (avérée ou présumée) ; une datation³⁹ ; un type⁴⁰ (pierre, papyrus, ostracon, parchemin, etc.). (2) Ce support est voué à la « conservation » — concept qui inclut l'objectif de transmission —, cadre duquel relève la problématique plus spécifique de l'inscription : chaque support impose ses caractéristiques, ses modes de segmentation, etc. ; un changement de support affectera généralement le texte. Il faut par ailleurs souligner que si, dans le cas de l'écrit, la conservation implique la mise en œuvre d'un système graphique (*i.e.* d'une sémiotique spécifique), ce dernier ne constitue pas une propriété du document dans la mesure où des textes recourant à des systèmes graphiques distincts peuvent être attestés sur un même document ; la variété des codes graphiques d'un document doit dès lors être déduite de la conjonction de ceux auxquels il est recouru par chacun des témoins. (3) Il est en outre nécessaire de préciser que tout document est susceptible de conserver « un ou plusieurs témoin(s) » ; en effet, chaque témoin étant la réalisation matérielle (de tout ou partie) d'un texte, la potentielle pluralité de témoins permet d'articuler la relation de plusieurs à plusieurs qu'entretiennent « texte(s) » et « document(s) » : un document est susceptible d'attester plusieurs textes et, réciproquement, un texte peut être transcrit sur plusieurs documents (ce qui équivaut à une multiplicité des témoins).

Définition 2. On nommera “texte” « tout agencement d'unités linguistiques par un (ou des) auteur(s) en vue de produire un message complet conservé par un (ou plusieurs) témoin(s). »

À la différence du document et du témoin, le texte est fondamentalement abstrait. (1) Tout texte repose sur un agencement (qui implique la mise en œuvre d'une syntagmatique) d'unités linguistiques ; ces dernières, appartenant à

38 Il n'est pas rare qu'un seul et même document soit connu sous plusieurs noms ; voir *e.g.*, dans le domaine égyptologique, le P. Bankes 1 (nom du propriétaire historique) qui est également connu sous le nom de P. BM EA 10023 (numéro d'inventaire du Service des antiquités égyptiennes du British Museum).

39 On se gardera de confondre la date du document (face matérielle) avec la date du (ou des) témoin(s) qui y sont transcrits, ainsi qu'avec la date de composition d'un texte *stricto sensu*.

40 Si l'on s'autorise à sortir du domaine des langues anciennes, ce type peut également être le papier, une bande magnétique (nous faisons donc une place à la notion de texte oral, nouvel objet philologique et matière de la linguistique de l'oral) ou un support numérique.

une ou plusieurs langue(s), délimitent les possibles univers de référence⁴¹ en relation avec les pôles extrinsèques du texte (l'auteur et le message). On touche alors aux dimensions extralinguistiques du texte. (2) Le rôle productif de l'auteur, qui s'inscrit dans une époque et une réalité sociale conditionnant (au moins partiellement) le type de texte produit, peut être envisagé en relation avec l'intention communicative du message, qui s'inscrit dans un genre et suppose un destinataire. Par ailleurs, (3) la notion de complétude possède une réelle importance : elle est directement liée à la volonté de l'auteur et renvoie à la cohérence globale d'une unité textuelle. Elle permet par conséquent de corréler la notion de texte à celle d'unité textuelle, entendue comme « le message complet d'un auteur, compte non tenu du support et du système linguistique mis en œuvre »⁴² [**définition 4**].

Définition 3. On nommera “témoin” « chaque transposition de tout ou partie d'un texte sur un document par un transcripteur. »

Le témoin est la face matérielle d'un texte, son actualisation sur un document. (1) La transposition de tout ou partie d'un texte sur un document est le fruit d'un transcripteur. Ce dernier peut-être identique à l'auteur ; le témoin est, dans ce cas, une source primaire du texte, ou témoin autographe. Si le transcripteur diffère de l'auteur, deux cas sont à considérer ; soit le transcripteur a travaillé avec l'auteur et, conceptuellement, il s'efface à la faveur de ce dernier, soit le transcripteur est indépendant de l'auteur et il représente une source secondaire donnant accès au texte de l'auteur. Cette double distinction permet de faire le départ entre les variantes « prétexte » de l'auteur⁴³ et les variantes de transmission découlant des différences de transposition entre les témoins secondaires d'un texte. (2) On remarquera que le transcripteur de plusieurs

41 Les problèmes complexes liés à la définition intralinguistique de la notion de « texte » s'amenuisent dès lors qu'on l'envisage de concert avec sa face matérielle (le témoin conservé sur un document), et en complémentarité avec les dimensions extralinguistiques (auteurs, message, etc.).

42 Ainsi, la traduction d'un texte produit un nouveau texte, quoique ce dernier relève de la même unité textuelle.

43 Le traitement de l'ensemble de ces variantes est, par exemple, pleinement intégré dans l'édition du texte de *Madame Bovary* sur le site de l'Université de Rouen (<http://bovary.univ-rouen.fr/>). Pour une approche théorique, voir C. Fuchs & al. (1982).

témoins peut endosser le rôle d'auteur. Il suffira, pour illustrer ce point, de songer au cas des anthologies : en transcrivant plusieurs témoins (partiels) de textes distincts, le transcripteur forme un nouveau texte dans la mesure où existent une nouvelle volonté d'auteur et une nouvelle idée de complétude⁴⁴. (3) Enfin, le transcripteur peut également endosser le rôle d'auteur lorsqu'il produit un nouveau texte indicé sur le texte transcrit. Le transcripteur glose ou commente alors le texte transposé sur un document et l'on entre dans le domaine des para- et métatextes.

Toutes ces définitions (qui ont pu sembler par trop abstraites) vont à présent être illustrées à travers la construction progressive du modèle d'encodage qui repose sur l'élément unificateur qu'est le dossier — la seule définition des trois entités n'étant pas suffisante pour rendre compte de l'ensemble des relations particulières qu'elles entretiennent entre elles. Aussi ces entités sont-elles intégrées dans un dossier, entendu comme l'ensemble des témoins (**déf. 3**) de chacun des textes (**déf. 2**) associés à une même unité textuelle (**déf. 4**), articulé par une « instance critique » (IC) dans le but de réaliser une « édition critique intégrale » (ECI), objectif final de la philologie numérique. C'est sur cette édition critique intégrale que porteront les analyses, quelles qu'elles soient ; il s'agit par conséquent du point de rencontre entre philologie et linguistique, si l'on se limite aux deux disciplines favorisées dans notre approche.

3. L'élaboration d'un dossier

Les dossiers ainsi formés sont, en fonction de leurs éléments constitutifs, plus ou moins complexes. Le modèle le plus simple est celui d'un dossier constitué d'un texte, représenté par un témoin unique inscrit sur un document (c'est par exemple le cas de nombreux textes épigraphiques et papyrologiques) ; un même texte pourra être représenté par plusieurs témoins (différentes copies d'une inscription ; différents manuscrits d'un texte ; etc.) ; un même dossier pourra, de la même manière, contenir plusieurs textes (dans le cas d'un texte traduit, par exemple) ; et un même document pourra porter la trace de témoins de plusieurs textes appartenant à différents dossiers (palimpseste ; etc.).

44 Il est entendu que chaque témoin partiel n'en reste pas moins un témoin du texte original, et appartient donc aux deux dossiers : celui de l'anthologie constituant un nouveau texte et celui du texte original.

En outre, de nombreux cas particuliers restent envisageables, comme la trace conservée d'un témoin dont le document aurait disparu (photographie ou estampage d'une inscription perdue ; édition d'un manuscrit perdu ; etc.), voire la conservation d'un document dont le témoin aurait disparu (pierre malade ; etc.).

3.1. Dossier contenant 1 texte, représenté par 1 témoin

Schématiquement, l'auteur (A) conçoit un texte, entité abstraite qui se matérialise sous la forme d'un témoin une fois transcrite sur un document (par le transcripteur originel [To]). La finalité est donc, à partir de la forme matérialisée du témoin, d'éditer un texte qui soit le plus proche possible du texte de l'auteur. Ce processus est le fait de l'instance critique (IC) et permet de dépasser l'édition diplomatique du témoin. La somme des différentes options retenues par l'instance critique constitue l'édition critique intégrale (ECI) sur laquelle porteront les multiples analyses.

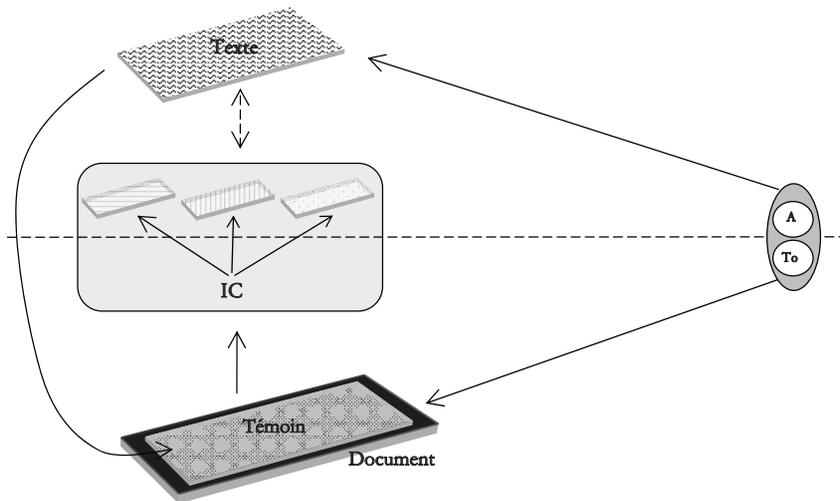


Figure 1. Témoin unique d'un texte unique

Le schéma simple du dossier comportant un unique témoin pour un unique texte est, dans les sources anciennes, largement représenté par les inscriptions et les papyri ou les ostraca documentaires qui transcrivent souvent des textes

confiés, par nature, à une copie unique (à titre d'exemples, lettre, inscription sur la base d'une statue, notes documentaires [comptes, inventaires, etc.]).

On en prendra pour exemple le P. *Cairo Zenon* I, 59034 (257^a)⁴⁵. Ce document porte la trace d'une lettre de Zoilos à Apollonios, ministre des finances, reçue le 13 février 257^a et enregistrée par Zénon, secrétaire particulier du ministre (le papyrus porte au verso la note d'enregistrement du secrétaire).

Dans ce premier schéma — comme dans ceux que nous examinerons ensuite —, un témoin ou un document peut lui-même présenter des particularités. Ainsi, l'unique témoin d'un texte peut n'être qu'une copie du document-témoin original. Il en va ainsi du sénatus-consulte et traité entre Rome et Astypalaia (105^a)⁴⁶. Ce sénatus-consulte a été inscrit sur une pierre, que l'on pouvait encore voir en Turquie à la fin du XVIII^e s. Elle fut copiée par J.-B.-G. d'Ansse de Villoison en 1785 dans le cimetière d'une église, à Astypalaia. La pierre, transférée à Smyrne, a disparu dans l'incendie de la ville en 1797, avant d'avoir été éditée. Aussi ce texte ne nous est-il plus connu que par le seul manuscrit de Villoison⁴⁷.

D'autre part, le document peut, comme nous l'avons déjà évoqué, comporter plusieurs témoins, qu'ils appartiennent ou non au même dossier. À fin d'illustration, le P. *Lit.Lond.* 165 (inv. 137)⁴⁸ porte, au recto, une doxographie médicale (dit « Anonyme de Londres ») et, au verso, a notamment été copiée une lettre d'Auguste au *κοινόν* d'Asie (33/32^a?). Le papyrus sera donc, comme document, intégré à deux dossiers distincts : la doxographie médicale, inscrite sur le recto de ce papyrus qui en constitue le seul témoin, suit ce premier schéma simple d'un dossier comprenant un témoin représenté par un document, tandis que la lettre d'Auguste fait partie d'un schéma plus complexe, dont nous verrons d'autres exemples dans la suite. En effet, la lettre originale était gravée sur une stèle qui, brisée de tous côtés, n'en conservait, au moment de sa découverte, que les fragments de 14 lignes⁴⁹. La pierre a disparu dans un

45 Voir W. Clarysse & K. Vandorpe (1995 : 78–85).

46 Voir R.K. Sherk (1969 : 16).

47 BnF, *Suppl. gr.* 930, f^o 78r–v (*Mémoire*, dans *MAIBL*, 47, 2 [1809], p. 285 ; Ét. Famerie 2006 : 125).

48 P. *Lit.Lond.* 165 (inv. 137), acquis par le British Museum en 1889 ; voir, pour la doxographie, W.H.S. Jones (1947), et, pour la lettre d'Auguste, R.K. Sherk (1969 : 57).

49 Le verso du papyrus porte la transcription de 35 lignes.

incendie en 1921 ou 1922, mais une trace de ce document a été conservée sous la forme d'une copie faite en 1871.

3.2. Dossier contenant 1 texte, représenté par $n > 1$ témoins

Ce schéma (cf. figure 2), qui regroupe les textes qui nous sont conservés par plusieurs témoins, est très largement représenté. L'exemple type, dans nos disciplines, en est le texte (généralement littéraire) que la tradition nous a transmis par plusieurs copies (papyrus, ostraca, codex ou manuscrits médiévaux).

L'œuvre de Pline (*Histoire naturelle*) en est une parfaite illustration : elle a joui d'un certain succès, avec pour conséquence un grand nombre de manuscrits (on en compte plus de 200)⁵⁰. Par ailleurs, la taille de l'œuvre a conduit à ce que de nombreux témoins n'en transmettent qu'une partie. De cette masse de copies de qualité inégale (de la copie fidèle au texte amendé et résumé), les critiques modernes ne retiennent pour l'établissement du texte que les meilleurs (en général les plus anciens).

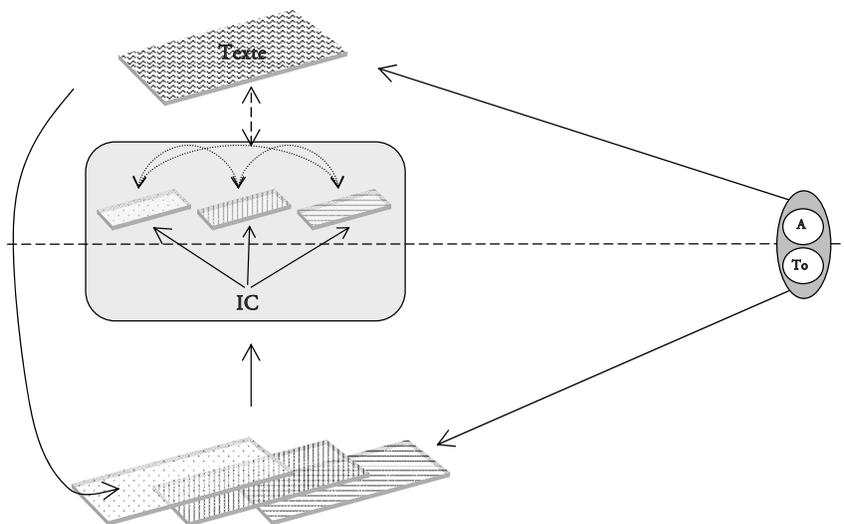


Figure 2. Plusieurs témoins d'un texte unique

On citera encore la lettre de M. Valerius Messala à Téos⁵¹ (qui était conservée sur le temple de Dionysos de ladite ville) ; la pierre est aujourd'hui perdue, et

⁵⁰ Voir A. Ernout (1950 : 20-35).

⁵¹ Voir R.K. Sherk (1969 : 34).

le texte de la lettre est connu par deux copies, dont dépendent toutes les éditions : la première de W. Sherard (1709), et la seconde de Ph. Le Bas (1843–1844)⁵². La copie de Sherard est connue par trois exemplaires⁵³ ; la copie de Le Bas est connue par l'estampage qu'il a réalisé de la pierre⁵⁴ et par l'édition du texte dans le *Voyage archéologique* de Le Bas & Waddington⁵⁵.

Comme on l'observe dans la figure 2, l'instance critique est le lieu d'un dialogue entre les leçons des différents témoins d'une part, et l'idée que l'on peut se faire de la complétude du message telle qu'elle fut envisagée par l'auteur au moment de produire le texte d'autre part. La philologie numérique possède cet avantage qu'elle ne demande pas d'opter définitivement pour l'une ou l'autre leçon (niveau ecdotique) conditionnant les possibilités d'interprétation (niveau herméneutique) ; elle autorise au contraire l'existence de leçons concurrentes (éventuellement hiérarchisées par l'instance critique) qui peuvent toutes être l'objet d'une analyse : le traitement de la variante est donc pleinement pris en compte dans le modèle proposé et vient féconder le niveau interprétatif.

En ce sens — quel que soit le schéma —, il faut souligner, au rang des particularités que peut présenter un document, la possible variabilité du témoin qu'il transmet. Ce peut être le cas pour un document dont la destination est modifiée (par exemple, une pierre portant la trace d'une inscription réemployée dans une construction, etc.), ou qu'altère une particularité physique. Il est ici opportun de rappeler le cas du *De magistratibus*⁵⁶ de Jean le Lydien.

Ce texte nous a été transmis par un manuscrit principal (dit *Caseolinus*) découvert par Viljoison à Constantinople en 1785, au cours de la chasse aux manuscrits dans laquelle il s'est lancé après sa découverte du *Venetus A*, et que Choiseul-Gouffier (ambassadeur auprès de la Sublime Porte) put récupérer. Le

52 Sur l'histoire de ces témoins, voir Ét. Famerie (2007).

53 Brit. Libr., *Ms. Add.* 10101, f° 116 (autographe), f° 172 (autographe ?) ; Brit. Libr., *Harl.* 7509, p. 30 (autographe ?).

54 L'estampage est aujourd'hui conservé dans les archives du Fonds Louis Robert de l'Académie des Inscriptions et Belles-Lettres.

55 Voir Ph. Le Bas & W.H. Waddington (1870 : III, 60).

56 Voir M. Dubuisson & J. Schamp (2006 : I, 2, DCCLXI-DCCLXXI).

manuscrit date probablement du x^e s., avec corrections du xvi^e ou xvii^e s.⁵⁷ On trouve la trace de quelques interventions d'une troisième main (peut-être celle qui a recopié le manuscrit au xviii^e). La singularité de ce manuscrit (O) est d'être recouvert de taches de vin, ce qui fit dire à Villoison : « *in cella uinaria potius eum quam in bibliothecam asseruatum uideri* ». Cette caractéristique a pour conséquence que le manuscrit évolue dans deux sens : certains endroits obscurs un temps sont devenus parfaitement lisibles par la suite, les taches s'étant éclaircies ; d'autres, jadis et naguère lisibles, sont devenus complètement indéchiffrables en raison du pâlissement de l'encre.

3.3. Dossier contenant $m > 1$ textes, représenté par $n > 1$ témoins

Sont rassemblées dans ce schéma les unités textuelles que la tradition nous a transmises dans différentes versions (états successifs ou concurrents d'une même œuvre, œuvre traduite, etc.).

Il en va ainsi des *Res gestae diui Augusti*⁵⁸. Le cas des hauts faits d'Auguste est en effet particulièrement complexe. D'après Suétone⁵⁹, Auguste aurait déposé quatre documents auprès des Vestales (un testament privé ; un texte d'ordres relatifs aux funérailles ; un résumé des œuvres accomplies ; et un texte sur la situation de l'Empire) ; Dion ajoute un cinquième document⁶⁰, qui aurait contenu des recommandations générales sur l'avenir. Les *Res Gestae* ne sont donc que l'un de ces *uolumina* ; Auguste voulait qu'elles fussent gravées sur des plaques de bronze et placées devant son Mausolée. Les copies romaines ont disparu. Ont été conservées des copies envoyées dans certaines provinces, à savoir :

- les copies d'Ankara (inscrites sur les murs du temple d'Auguste et de Rome), conservant le texte latin sur les murs intérieurs du pronaos et une traduction grecque sur les parois extérieures ;
- la copie d'Antioche, conservant 270 fragments portant le texte d'une copie latine ;
- la copie d'Apollonia, contenant des fragments d'une traduction grecque.

57 Ces corrections sont exclusivement de type orthographique (fautes, ajout d'accents et esprits).

58 Voir J. Scheid (2007).

59 SUÉT., *Aug.*, 101, 1 et 6.

60 DION CASS., *LVI*, 33, 2-6.

On dispose ainsi de 4 copies, qui transmettent le même texte original (avec seulement quelques variantes imputables aux habitudes des scribes : chiffres romains ou en toutes lettres ; abréviations ; ordre des mots ; oubli ; iotacisme ; etc.). Le stemma, simple, est le suivant⁶¹ :

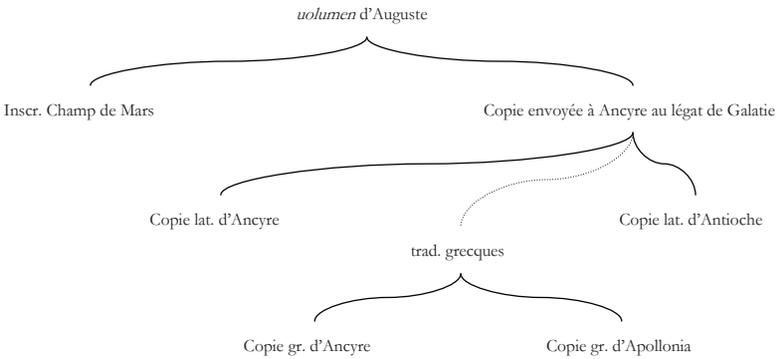


Figure 3. Stemma des *Res gestae diui Augusti*

On constatera, dans le schéma qui suit, un dialogue entre les différents textes au niveau de l'instance critique :

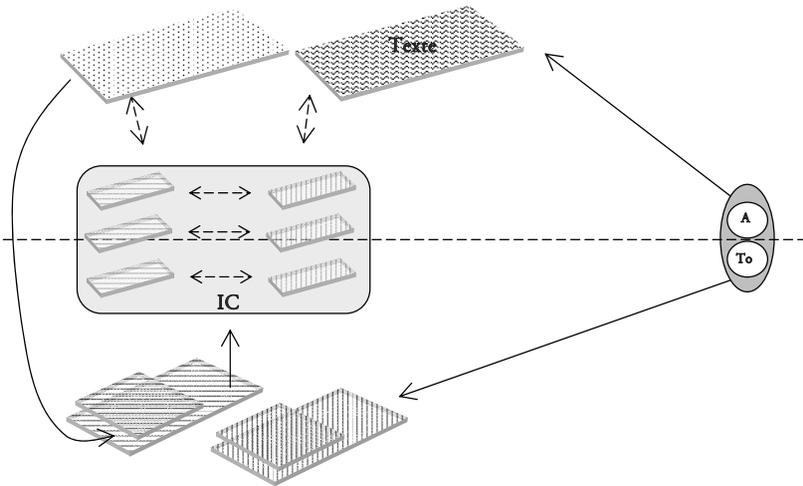


Figure 4. Unité textuelle composée de deux textes conservés par plusieurs témoins

61 Cf. J. Scheid (2007 : xvi). Les deux traductions proviennent d'une même source ; s'il n'est pas possible de déterminer où elles ont été réalisées avec certitude, il est vraisemblable qu'elles aient été faites par un Grec, peut-être du secrétariat du légat de Galatie.

Longtemps, les éditeurs des *Res gestae* ont repris le texte d'Ancyre en le truffant de signes critiques pour rendre compte des autres inscriptions, au détriment d'une édition critique et synthétique. La présence au sein d'un même dossier de deux textes distincts — l'un en latin et l'autre en grec — pose directement la question de la granularité des possibles procédures d'alignement entre les témoins d'un dossier donné dans l'instance critique. La position méthodologique la plus satisfaisante en ce domaine nous paraît être la suivante : tous les témoins d'un même texte, dans la mesure où ils sont rédigés dans une langue commune (et correspondent à une volonté de complétude identique), peuvent être alignés au niveau de l'unité minimale qu'est le mot (quelles que soient les difficultés définitionnelles liées à cette notion) ; en revanche, l'alignement entre les témoins de deux (ou plusieurs) textes distincts dans un même dossier ne peut être pertinente qu'au niveau de la proposition — le changement de code linguistique (ou la modification de l'intention communicative, dans le cas d'éditions successives dues à l'auteur par exemple) n'autorisant pas une granularité plus fine.

Par conséquent, on peut préciser les opérations conceptuelles relevant de l'instance critique. Elles sont globalement de deux types : l'intégration des informations ecdotiques pertinentes découlant d'un travail critique sur l'ensemble des témoins d'un dossier (compte non tenu des textes concernés) et, parallèlement, l'alignement de tous les témoins à deux niveaux, respectivement celui du mot pour les témoins relevant d'un même texte et celui de la proposition dans les autres cas. Cette dernière remarque nous conduit aux conclusions de ce travail préliminaire à la modélisation de la relation texte(s)-document(s).

4. Conclusion

Une fois les opérations de l'instance critique réalisées, toutes les données afférentes à l'analyse d'un texte (quelles que soient la nature et la portée de celle-ci, cf. *infra*) deviennent disponibles dans ce que l'on a nommé « édition critique intégrale » (ECI). Cette dernière dépasse la pratique philologique traditionnelle de collation des manuscrits qui a pour objectif la reconstruction du texte originel. En effet, le modèle proposé pour la philologie numérique permet, à partir de l'ECI, d'accéder à l'ensemble des variations significatives : variantes d'auteur (prétexte), variantes de transmission dues à la tradition manuscrite, variantes d'édition fruit du travail critique, etc. L'établissement d'un dossier

dans toute sa complexité (qui suppose l'articulation des textes, témoins et documents qu'il contient) permet dès lors de ne pas préjuger des balisages possibles lors des analyses et, par conséquent, de ne pas limiter d'avance — voire de réduire à néant — certaines exploitations postérieures. L'herméneutique, point d'orgue de la philologie, s'en trouverait améliorée d'autant.

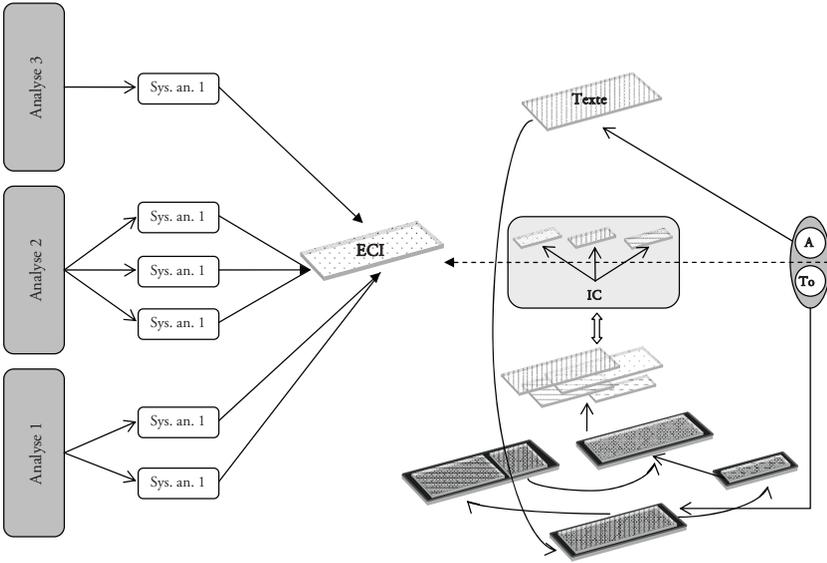


Figure 5. Schéma général

Comme on a eu l'occasion de le souligner à différentes reprises, c'est sur l'ECI que porteront les analyses dans la mesure où elle permet d'accéder à tous les types de données utiles à l'étude d'un texte. Ainsi que le montre la figure 5, chaque analyse (qu'elle soit linguistique, diplomatique, thématique, etc.) devrait générer un système d'annotation spécifique et, subséquemment, chaque chercheur pourra — en fonction de ses besoins — définir son jeu d'annotation personnel⁶². Pour se limiter au cadre de l'analyse langagière, l'application généralisée de ce principe conduira à une linguistique qui fait droit à tous les niveaux de complexité de son objet : du mot au texte au cœur d'un dossier,

62 Cette idée est directement empruntée aux travaux que N. Mazziotta a menés dans le cadre du Projet Khartès (Université de Liège). On mentionnera plus particulièrement le logiciel *NotaBene* qu'il développe actuellement pour permettre l'annotation linguistique multiple. Sur ce logiciel, voir N. Mazziotta (à paraître).

puis du texte au corpus par la combinaison de dossiers en fonction de critères objectivables, car s'appuyant sur un (ou plusieurs) élément(s) définitionnel(s) des dossiers.

Les réflexions proposées dans cette contribution ne constituent, cela va sans dire, qu'une série de principes préliminaires qui mériteraient d'être précisés sur bien des points. Cependant, la cohérence du modèle nous semble suffisante pour mener, à terme, à une spécification qui permette une implémentation effective de la relation texte(s)-document(s) et, partant, conduise à de nouvelles pratiques en matière de philologie numérique.

Bibliographie

- J.-M. ADAM (2000), *Textes et genres de discours*, Paris.
- (2004 [1999]), *Linguistique textuelle. Des genres de discours aux textes*, Paris.
- P. ANDRIES (2008), *Unicode 5.0 en pratique. Codage des caractères et internationalisation des logiciels et des documents*, Paris.
- J. SCHEID (2007), *Res gestae diui Augusti*, Paris.
- S. BADIR (1998), *La notion de texte chez Hjemslev*, dans *Texto !* [http://www.revue-texto.net/Dialogues/Debat_Hjemslev/Badir_Notion.html].
- E. BARBOTIN [dir.] (1975), *Qu'est-ce qu'un texte ? Éléments pour un herméneutique*, Paris.
- R. BARTHES (1972), *s.v. Texte*, dans *Encyclopaedia Universalis*, t. 15.
- L. BURNARD & S. BAUMAN [éd.] (2008), *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, Oxford [<http://www.ehess.fr/fr/enseignement/enseignements/2008/ue/3683/>].
- W. CLARYSSE & K. VANDORPE (1995), *Zénon, un homme d'affaires grec à l'ombre des pyramides*, Louvain.
- J. DENOZ (1978), *L'ordinateur et le latin. Techniques et méthodes*, dans *Revue de l'organisation internationale pour l'étude des langues anciennes par ordinateur* 4, p. 1-36.
- (2007), *Langues classiques et informatique : premiers travaux*, dans Fr. CITTI & T. DEL VECCHIO [éd.], *Fom Manuscript to Digital Text, Proceedings of the Colloquium (Bologna, June 12th 2003)*, Rome.
- M. DUBUISSON & J. SCHAMP (2006), *Jean le Lydien. Des magistratures de l'État romain*, 3 vol., Paris, Belles Lettres.
- A. ERNOUT (1950), *Pline l'Ancien. Histoire naturelle. Livre 1*, texte établi, traduit et commenté par J. Beaujeu, introduction par A. Ernout, Paris.

- Ét. FAMERIE (2006), *J.-B.-G. d'Ansse de Villosion. De l'Hellade à la Grèce. Voyage en Grèce et au Levant (1784-1786)*, Hildesheim (= *Altertumswissenschaftliche Texte und Studien* 40).
- (2007), *Papiers de Sherard, copies de Hochepped, schedae de Duker : contribution à l'histoire des copies manuscrites des inscriptions de Teos*, dans *Chiron* 37, p. 65–88.
- C. FUCHS, A. GRESILLON, J.-L. LEBRAVE, J. PEYARD & J. REY-DEBOVE (1982), *La genèse du texte : les modèles linguistiques*, préface d'A. Culioli, Paris.
- B. HABERT, C. FABRE & F. ISSAC (1998), *De l'écrit au numérique : constituer, normaliser et exploiter les corpus électroniques*, Paris.
- L. HJELMSLEV (1968-1971), *Prolégomènes à une théorie du langage*, Paris.
- H. ISENBERG (1970), *Der Begriff Text in der Sprachtheorie*, dans *ASG-Bericht* 8, p. 1-21.
- W.H.S. JONES (1947), *The Medical Writings of Anonymus Londinensis*, Cambridge.
- R. KYHENG (2005), *Hjelmslev et le concept de texte en linguistique*, dans *Texto* 10/3 [http://www.revue-texto.net/Inedits/Kyheng/Kyheng_Hjelmslev.html].
- Ph. LE BAS & W.H. WADDINGTON (1870), *Voyage archéologique en Grèce et en Asie Mineure*, 3 vol., Paris.
- N. MAZZIOTTA (en préparation), *Logiciel NotaBene pour l'annotation linguistique. Annotations et conceptualisations multiples*, dans V. DUMONT & C. LEJEUNE [éd.], *Recherches qualitatives*, Numéro Spécial : *Logiciels pour l'analyse qualitative*, Trois-Rivières
- S. MELLET & Ét. BRUNET, *Hyperbase. Manuel de référence pour la base latine*, distribué par l'UMR 6039 (*Bases, corpus et langage*).
- Fr. RASTIER (1996), *Pour une sémantique des textes. Questions d'épistémologie*, dans Fr. Rastier [éd.], *Sens et textes*, Paris, p. 9-35.
- (2000), *Philologie numérique*, dans J.-E. TYVAERT [éd.] (2000), p. 117-149.
- (2001), *Arts et sciences du texte*, Paris.
- S. ROSMORDUC, St. POLIS & J. WINAND (2008), Ramses. *A New Research Tool in Philology and Linguistics*, dans *Informatique et Égyptologie – Proceedings*, p. 133-142.
- F. DE SAUSSURE (1972 [1916]), *Cours de linguistique générale*, édition critique préparée par Tullio De Mauro, Paris.
- R.K. SHERK (1969), *Roman Documents from the Greek East. Senatus Consulta and Epistulae to the Age of Augustus*, Baltimore, Johns Hopkins Press.
- O. SOUTET (2001), *Linguistique*, 3^e édition, Paris.
- J.-E. TYVAERT [éd.] (2000), *Philologie électronique et assistance à l'interprétation des textes*, dans *Recherches en linguistique et psychologie cognitive* 16.

- J.-M. VIPREY (2006), *Quelle place pour les sciences des textes dans l'Analyse de Discours ?*, dans *Semen 21 (Catégories pour l'analyse du discours politique)*, mis en ligne le 28 avril 2007 [<http://semen.revues.org/document1995.html>] .
- G. WILLIAMS [éd.] (2005), *La linguistique de corpus*, Rennes.
- J. WINAND, St. POLIS & S. ROSMORDUC (2009 – sous presse), Ramses. *An Annotated Corpus of Late Egyptian*, dans *Actes du X^e Congrès International des Égyptologues*, éd. par P. Kousoulis, Leuven, 9 p. (= *OLA*).
- Th.N. WINTER (1999), *Roberto Busa, S.J., and The Invention of the Machine-Generated Concordance*, dans *Classical Bulletin* 75, p. 3-21.

