Édition numérique de documents textuels

Vers un modèle d'infrastructure pour la critique textuelle à partir des méthodes, expériences et prototypes développés à l'ILC de Pise

Andrea Bozzi – ILC/CNR – Pise

[Power Point prima parte (Parigi_16_01_2014)]

[Slide 1. Titolo]

Introduction

Le thème de l'édition électronique de documents numériques est de plus en plus débattue en étroite relation avec les campagnes de numérisation des fonds d'archives et de bibliothèques qui ont eu lieu au cours des dernières années. Cela a été rendu possible grâce au financement public ou aux subventions de grandes entreprises privées, en obtenant des résultats quantitativement très pertinents et qualitativement excellents. Il est inutile de mentionner *Gallica*, que nous connaissons tous, ou les initiatives de la *Bibliothèque Apostolique Vaticane* à Rome, qui a annoncé la mise à disposition des chercheurs d'une masse de données numériques d'environ 43 pétaoctets (PB) : cela signifie que dans quelques années, des millions de pages seront accessibles en ligne aux chercheurs. Je n'ai pas l'intention de parler d'*Europeana*, une autre initiative bien connue qui confirme, s'il en était besoin, une tendance maintenant inéluctable pour le futur accès à la culture "livresque". Chaque analyse pourrait, par conséquent, paraître incomplète parce que cet accroissement se produit à un rythme quotidien : ce qui est fiché aujourd'hui est déjà incomplet le lendemain.

Comme indiqué en introduction, les questions à examiner lors de ce séminaire pourraient être nombreuses. Mais je n'ai pas le don de l'omniscience et je dois donc limiter mes observations à quelques questions dont je crois avoir acquis une certaine expérience et expertise au cours de ma carrière de chercheur qui, du point de vue bureaucratique, s'est terminée le 31 mai, date de mon départ à la retraite.

Par conséquent, je vais essayer d'énumérer les sujets dont je vais traiter. Certains seront traités plus superficiellement parce que je les connais moins, même si je pense qu'ils sont très importants, d'autres par contre seront examinés beaucoup plus en détail. Parmi eux, je vais souligner les aspects de la méthode et, le cas échéant, les expériences directes réalisées avec des outils qui sont fondés sur ces méthodes.

Le but ultime de cette présentation est très ambitieux : je voudrais soumettre à votre évaluation et votre opinion un "modèle" de texte numérique et d'édition numérique du texte (manuscrit et imprimé) que les essais expérimentaux effectués à l'ILC laissent entrevoir.

[Slide 2. Esempi di sistemi Collate/Anastasia - De Montfort University e University of Saskatchewan ; TUSTEP TUebingen System of Text Processing Programs - Tuebingen University ; CTE (Classical Text Editor) ; EDMAC Critical Edition Typesetting ; CET (Critical Edition Typesetter), ecc.]

De nombreux centres de traitement électronique pour les sciences humaines ont créé des logiciels à la demande des chercheurs de disciplines diverses dont les objectifs peuvent diverger. L'idée de "Pisa", cependant, vise à atteindre une « infrastructure » éditoriale flexible, organisée sous la forme d'une application Web à modules intégrés, afin de répondre à une variété d'utilisations spécialisées. Dans le même temps, elle doit répondre aux conditions requises pour la formation universitaire des jeunes philologues numériques.

[Slide 3 : elementi di base per la Digital Philology]

Les exemples montrent que les éléments de base de tous les projets mentionnés sont essentiellement au nombre de quatre :

1. numérisation de sources primaires ;
2. métadonnées descriptives du texte et des informations qui ont été ajoutés par des outils automatiques ou semi-automatiques ;
3. préparation de l'édition critique ;
4. production des résultats sur papier ou sur Internet, avec des systèmes de navigation dédiés.

En un mot, ces éléments représentent la philologie numérique selon un modèle simplifié dans le graphique ci-dessous :

[Slide 4 : modello della filologia digitale]

Les éléments centraux de cette présentation seront par conséquent :

[Slide 5 : lista dei contenuti della presentazione]

- La question de la **relation entre les sources primaires (*primary sources*) et les sources secondaires** (*annotaded sources*, au sens où les sources secondaires correspondantes aux sources primaires sont enrichies des informations produites par des outils automatisés, semi-automatiques ou entièrement manuels). Cet aspect est étroitement lié à celui des métadonnées descriptives et des standards.
- La question de la **traduction des sources primaires**, considérée dans le modèle comme une forme particulière de l'information secondaire.
- La question des apparats critiques, c'est-à-dire de la production d'éditions critiques de documents numériques, soit anciens soit modernes, en particulier des manuscrits.

Tous ces éléments sont interdépendants : plus on ajoute d'informations sur les données, plus importantes seront les possibilités expressives de ces mêmes données pour répondre aux besoins d'une multiplicité d'utilisateurs. La granularité de l'information et la « ré-utilisabilité » des objets numériques jouent un rôle clé dans le progrès technologique de toutes les sciences humaines, pour la recherche comme pour la formation de nouveaux professionnels dans le domaine des études littéraires et de la philologie. Entre autres choses, ce thème est central pour préparer une éventuelle proposition de collaboration entre l'Europe et les États-Unis, entre des centres tels que l'ILC-CNR, le Labex Obvil (Italie-France) et l'Hyperstudio (MIT, Boston). Le thème de la formation est également au centre de toutes les activités récentes du désormais européen Gregory Crane, qui a déménagé de l'Université Tufts de Boston à Leipzig pour au moins cinq ans, grâce à un *Von*

Humboldt Stiftung étroitement lié aux activités de l'informatique des sciences humaines et de la philologie numérique.

[Slide 6 : lista dei contenuti di cui non parlerò]

Je ne traiterai pas de thèmes, bien que très importants, tels que :

— l'édition numérique de documents musicaux avec une possible corrélation entre le texte de la partition et les fichiers audio correspondants ;

— l'édition et/ou le commentaire et l'annotation des fichiers vidéo.

Je vais vous mentionner brièvement seulement ces deux questions parce que le « modèle de Pise » n'exclut pas de pouvoir gérer ce type de fichier.

[Slide 7 : Lista dei contenuti della conferenza : evidenziazione del primo punto]

[Slide 8 : Fonti primarie e metadati]

1. Sources primaires, sources secondaires, métadonnées

Il y a au moins deux types de métadonnées :

- Les métadonnées utiles à la gestion d'un document. En Italie, ils sont appelés MAG (Métadonnées Administratives Gestionnelles) : il s'agit d'un standard pour collecter des métadonnées liées aux objets créés dans des projets de numérisation. Les marqueurs appartenant à ce type définissent des aspects tels que la date de création de l'objet numérique, les caractéristiques techniques de conversion analogique-numérique, l'éventuelle opération de retouche et d'*enhancement* effectuée au cours du processus de numérisation, etc. Du point de vue du chercheur, ces données sont relativement peu importantes, car ces dernières ne concernent pas le contenu même de l'objet numérique plus intéressant à ses yeux.
- Il existe une couche supplémentaire d'informations que l'objet numérique transmet et qui est très difficile à quantifier parce qu'il y a une relation bi-univoque entre l'objet de l'enquête qu'on entend mener sur l'objet numérique et les informations qu'il contient, capables de répondre aux besoins spécifiques de l'étude. En d'autres termes, les informations explicites ou implicites décrivant un objet numérique sont de nature variée, et leur marquage dépend de l'objectif à atteindre en utilisant cet objet. Pour donner un exemple : l'image numérique de la Tour Eiffel contient plusieurs informations qui peuvent être représentées par des métadonnées de différents types, selon la manière dont on souhaite la présenter : une attraction touristique, le symbole de Paris, un spécimen du monument de fer, une œuvre artistique et architecturale, etc. *C'est pourquoi l'objet numérique, comme n'importe quel objet de la réalité, est polyvalent du point de vue de l'information et c'est souvent le contexte dans lequel il se trouve qui aide à sélectionner une partie de son potentiel d'information et à éliminer d'autres informations.*

La décontextualisation d'un objet numérique poserait de graves difficultés dans la sélection des métadonnées qui peuvent représenter les fonctions de communication "universelles" qui sont selon moi des éléments trompeurs.

Cela étant dit, je peux maintenant présenter mes prochaines observations liées au domaine des textes "numériques" dans une perspective seulement d'enquête philologique, sans oublier toutefois, l'aspect de l'enseignement et de la formation, que je n'ai pas l'intention de séparer de celui de la recherche.

Pour moi, ces deux aspects (*scholarly* e *educational*) sont indissociables.

[Slide 9 : suddivisione dell'oggetto digitale di tipo "testo" per l'attribuzione di metadati per la ricerca filologica]

Un document numérique de type "texte" (écrit, ou la transcription d'un texte parlé), est caractérisée par une quantité considérable d'informations qui peuvent être classées, sans prétendre tout à fait à l'exhaustivité, au moins dans les catégories suivantes :

- *Éléments extra-textuels ;*
- *Éléments para-textuels ;*
- *Le texte lui-même.*

[Slide 10]

1.1 Éléments extra-textuels

Un texte véhicule au moins deux types différents d'information : la première est représentée par des "individus typographiques", extérieurs. Dans certaines publications antérieures, je les avais définis d'une façon peut-être trop simpliste, avec le terme d' « éléments extra-textuels » :

- la numérotation de la page ou de la feuille ;
- la numérotation des chapitres et des paragraphes ;
- le *layout* correspondant à la disposition de la zone d'écriture sur la page (le nombre de colonnes, le nombre de lignes) ;
- la numérotation des vers d'un texte poétique ;
- les titres courants, très fréquents dans les ouvrages encyclopédiques ou les lexiques anciens ;
- tous les éléments qui caractérisent la surface d'écriture sur laquelle le texte est transmis jouent un rôle subordonné et, dans de nombreux cas, ils ne dépendent pas d'indications originales, mais ont été insérées a posteriori à des fins éditoriales ou rédactionnelles.

[Slide 11]

1.2 Éléments para-textuels

Il y a un deuxième type d'informations que j'avais défini comme "éléments para-textuels", fixés à un niveau de communication certainement supérieur aux "individus" du type précédent.

Font partie de ce type des "individus" qui contribuent à augmenter, spécifier, détailler le contenu informatif du texte comme par exemple les annotations, les interventions marginales, les notes, les initiales enluminées, les miniatures ou les illustrations des manuscrits ou des livres imprimés anciens, les légendes, etc. Citons encore les données enregistrées dans un apparat critique, les variantes et passages alignés qui sont autant d'éléments para-textuels. Je discuterai de cet aspect plus loin (voir paragraphe 4), car il fait partie d'un plan d'accords scientifiques et techniques avec le Labex OBVIL dirigé par mon collègue, le professeur Didier Alexandre.

Les métadonnées relatives à cette typologie d'informations sont extrêmement importantes car dans de nombreux cas, comme par exemple pour les illustrations ou pour les graphiques des textes de l'histoire des sciences, elles ont une valeur communicative comparable à celle du texte lui-même. Ne sont pas rares les cas où les erreurs du texte attribuables à un imprimeur sont en fait amendables en recourant à l'illustration qui est au contraire correcte.

[Slide 12]

1.3 Le texte

Les éléments précédents doivent donc être considérés à un autre niveau que celui dans lequel il se trouve cependant : le “texte lui-même”. Il s’agit de la principale source d’information et il est composé de différents sous-types :

- le texte linguistique considéré comme une séquence de *tokens* (signifiants) ;
- le texte conçu comme un contenu métalinguistique exprimé par les modalités, souvent de type graphique, avec lesquelles les mots apparaissent sur la page ;
- enfin, le texte conçu comme un contenu sémantique, exprimé par des éléments linguistiques enchaînés dans des phrases (signifiés) ; par la traduction (comme nous le verrons dans la section 2 « La traduction des sources primaires »).

Voyons plus en détail les “individus” qui caractérisent chacun de ces trois ensembles :

[Slide 13]

La séquence de *tokens*. Je n’ai pas l’intention d’insister sur cet aspect, bien connu par ceux d’entre vous qui ont l’expérience de l’analyse de texte informatisé. La seule chose sur laquelle je voudrais mettre l’accent est étroitement liée à ce que l’on doit entendre par *token* et, en conséquence, ce que doit faire un analyseur (*parser*) automatique, capable de distinguer les *tokens* d’un texte pour attribuer la classification « lexèmes » aux mots, « diacritiques » aux symboles de ponctuation, etc. Même la ponctuation, qui semble posséder une faible valeur heuristique et se soumettre à des interprétations subjectives mineures, peut en fait être chargée d’un potentiel sémantique significatif. Il suffit de penser à l’introduction de la ponctuation par les éditeurs modernes de manuscrits anciens où elle était absente : différents emplacements dans le texte, ou éliminations des signes de ponctuation contribuent au changement du sens d’une phrase entière.

Même l’espace utilisé par l’analyseur (*parser*) pour reconnaître chaque *token* joue un rôle clé dans la sémantique d’un texte : que se produit-il si les anciens manuscrits écrits en *scriptio continua* sont interprétés de façon arbitraire, en séparant les chaînes de caractères pour former plusieurs *tokens* différents ? La marge d’arbitraire est très élevée et les compétences linguistiques et philologiques ne peuvent pas toujours résoudre les doutes. Dans le code Hamilton de les “*Novelle*" de Giovanni Boccaccio, par exemple, il y a beaucoup de cas de ce genre, et des incertitudes demeurent pour la segmentation de chaînes telle que « *cheglisidonasse ». On peut* transcrire « *che gli si donasse »* ou « *ch’egli si donasse ».* Ces deux segmentations du texte en unités lexicales ont des significations complètement différentes en italien.

[Slide 14]

Les éléments métalinguistiques. Dans cette section, nous faisons face à des situations très communes dans les manuscrits autographes d’auteurs modernes et contemporains. L’utilisation du soulignement, d’annotations colorées, de ratures ou d’autres interventions de ce type a une signification pas toujours facile à restituer par un éditeur critique, mais souvent indispensable pour l’étude de l’œuvre qui apparaît, dans la version numérique, dans toute sa complexité. Les phénomènes sont très fréquents même dans les manuscrits anciens et dans les anciennes éditions imprimées où on trouve des notes inscrites par les lecteurs, ainsi que des interventions des auteurs eux-mêmes, quand ils ont décidé d’apporter des modifications après la publication des volumes.

Quelles sont les méthodes à mobiliser pour ne pas perdre les informations importantes qui sont transmises par ces éléments métalinguistiques ? Il existe une norme pour déterminer si et dans quelle mesure il est nécessaire d'intervenir avec un critère de marquage, c'est-à-dire avec l'introduction de métadonnées grâce auxquelles cette information est récupérable et utilisable par un système d'interrogation.

En effet, si les spécialistes ont aujourd'hui à leur disposition des outils de marquage très efficaces et complets, il n'existe pour autant pas de critères généraux à suivre pour créer sur le Web des archives de textes numériques à des fins d'enquête linguistique et philologique. Ces dernières années, compte tenu du succès et de la diffusion des langages de codage tels que le standard TEI, plusieurs manuels ont été publiés en Italie, centrés sur le thème de l'introduction de métadonnées descriptives à des niveaux de granularité très fins. Certains par exemple, accordent une attention particulière aux aspects paléographiques qui permettent de distinguer les sections d'un manuscrit médiéval et documentent la présence de différents écrivains. D'autres poussent l'analyse jusqu'à décrire en détail l'état de conservation du support papier qui limite la lecture et l'interprétation du texte. On pourrait ajouter une grande quantité d'autres exemples, **mais une question s'impose : tout cela est-il vraiment utile ?** Ne courons-nous pas le risque de charger un spécialiste de trop d'efforts et d'un travail lourd, uniquement parce que la technologie numérique permet, contrairement à la technologie analogique, de reproduire et de marquer les nombreux éléments présents dans le texte original ?

[Slide 15]

Les significations : la lemmatisation. Afin d'étayer cette hypothèse, je voudrais mentionner aussi le cas de la lemmatisation qui, d'une manière seulement très apparente, pourrait sembler le moins susceptible de subjectivisme et par conséquent, plus sûr et plus valable. Le lemme, auquel sont associées toutes les formes qui lui appartiennent morphologiquement et pas de façon ambiguë, n'est pas une entité toujours si évidente à déterminer. Lorsqu'on étudie des textes anciens écrits dans des langues que nous ne maîtrisons pas comme ceux qui les parlaient (latin, grec, sanskrit, vieux slave, etc.), le lemme devient souvent un "artifice" utile, en particulier lors de la construction de lexiques à l'aide d'un ordinateur, mais pas toujours acceptable. On peut mentionner, pour la langue latine, les formes nominales des verbes représentées par les participes présents ou passés. Elles sont morphologiquement des formes verbales, mais selon le contexte, la période ou l'œuvre, elles doivent être considérées comme des substantifs. ABSOLUTUM est un substantif qui ne peut pas avoir comme POS (*Part of Speech*) la métadonné qui le classifie comme participe passé de ABSOLVO, parce que dans certains domaines il "est" un substantif à tous les égards.

Souvent, les aspects de la lemmatisation se croisent avec ceux de la tokenization. Du point de vue instrumental et opérationnel, nous considérons les *tokens* comme des unités minimales produites par l'analyseur (ou tokenizer), mais nous savons tous qu'un *token* peut correspondre à un mot pour certains, mais pour d'autres à plusieurs mots qui "doivent" être concatenés pour former une unité linguistique autonome. Par exemple : « *erba di San Giovanni »* pourrait être considéré comme un ensemble de 4 *tokens*, alors qu'en réalité, il s'agit d'un élément linguistique unique qui a comme synonyme le mot "iperico" (le millepertuis ou *Hypericum perforatum*).

[Slide 16]

Cette diapositive illustre bien ce propos**.** Elle montre la page du manuscrit inédit de Saussure vue auparavant. Il apparaît que les modalités d'annotation et par conséquent la création de

métadonnées dépendront des objectifs de la recherche qu'on souhaite poursuivre sur ce texte. Un spécialiste de la linguistique historique cherchera à mettre en évidence des aspects différents de ceux qui retiennent un sociolinguiste ou un théoricien du langage.

Sans entrer dans le détail, je veux rappeler ici que les objets numériques semblent fortement prédisposés à un relativisme de l'interprétation qui va au-delà du contexte dans lequel ils sont placés. Le simple fait qu'ils soient disponibles dans une dimension potentiellement universelle, les expose à une classification ou taxonomie difficile : les "visites" avec lesquelles ils sont observés dans le macrocosme du réseau sont innombrables et parfois imprévisibles.

[Slide 17]

Permettez-moi de conclure cette partie de mon argumentation en disant que :

- il n'existe pas un critère univoque selon lequel un objet numérique, texte ou autre, peut être défini avec une seule métadonnée descriptive, puisque la description dépend du sujet qui "interprète" et "classe" cette donnée. Contrairement à l'approche philosophique qui attribue à un objet une validité ontologique unique dans l'univers du monde réel, la virtualité de la technologie numérique relativise le concept de l'ontologie : un objet est ce qu'il est par rapport à l'observateur dans un contexte déterminé ;
- la classification d'un objet numérique est par conséquent, une méthode itérative qui accumule les métadonnées de différents types selon le point d'observation, le contexte dans lequel l'objet est inséré et la compétence ou les objectifs de l'observateur ;

[Slide 18]

- l'aspect collaboratif du Web rend possible (mais pas obligatoire) cette accumulation et représente un outil puissant pour l'enrichissement des objets numériques et leur utilisation par différents types d'utilisateurs ;
- dans une perspective didactique et de formation de nouveaux concepts, les étudiants peuvent faire partie intégrante de cette procédure de classification et d'attribution de métadonnées descriptives.

Je voudrais maintenant représenter graphiquement la procédure d'attribution des métadonnées d'un objet numérique qui passe du statut de source primaire à celui de source secondaire, chargée avec des éléments descriptifs qui en accroissent la valeur.

[Slide 19 : grafico]

Comme on le voit dans la figure, nous pouvons imaginer que l'objet numérique primaire a-sémantique, non annoté, est placé au centre d'un cercle d'où, à mesure que l'objet est analysé, partent des rayons qui forment des sections correspondant aux types d'annotation et de métadonnées qui s'y rapportent. Chaque section, comme nous le disions ci-dessus, se réfère maintenant aux métadonnées linguistiques, philologiques, ecdotiques, historiques, graphiques, sémantiques-lexicales, etc.

Je trouve que ce paradigme interprétatif est efficace pour exprimer la façon dont je vois les différentes modalités de description d'un objet numérique. Il est aussi utile pour simplifier l'accès aux données grâce à une application Web. En fait, si les métadonnées sont représentées par des langages standardisés, des interfaces appropriées permettent l'inclusion ou l'exclusion des

sections (c'est-à-dire des métadonnées) qui sont considérées comme utiles ou inutiles pour les besoins d'une navigation spécifique du texte et de son contenu.

Il s'ensuit que :

[Slide 20]

- les métadonnées sont divisées en sections : extra-textuelles, para-textuelles et textuelles ;
- chaque section contient autant de sous-sections que celles repérées et balisées par chaque chercheur dans le texte analysé ;
- le bord supérieur des sections n'est pas défini : il est très peu probable qu'un objet numérique soit décrit dans « toutes » ses composantes : à mon avis cette descrition est subjective et très dépendante de la sensibilité, de la compétence et de l'expérience de l'observateur/annotateur ;
- dans les phases d'interrogation, les sections constituées de différents types de métadonnées peuvent être choisies de manière sélective afin que l'application accède aux données pour répondre aux besoins de consultation par les différents types d'utilisateurs (par exemple, les utilisateurs experts ou débutants).

Selon ce modèle, il n'est pas important de tendre à l'exhaustivité de la description d'un objet numérique, que je considère impossible ; mais il est stratégique de proposer aux diverses catégories d'usagers la possibilité de savoir quels types d'annotation caractérisent un objet et de sélectionner les métadonnées qui sont importantes pour leurs besoins de navigation[1].

Pour ne pas rester sur un plan théorique et pour expliciter, avec des exemples très simples, quelles sont les implications pratiques de ce que j'ai dit, je voudrais montrer quelques cas qui, à mon avis, sont emblématiques.

[Slides 21, 22, 23, 24, 25, 26]

2. La traduction des sources primaires

[Slide 27 : lista dei contenuti con evidenziato il secondo punto]

Comme nous l'avons vu dans les dernières diapositives, j'ai voulu délibérément montrer des exemples dans lesquels apparaissent un texte et sa traduction : dans notre modèle, la traduction joue un rôle important et est inséré entre les informations « sémantiques » du texte lui-même. Il a une correspondance avec la composante lexico-sémantique parce que la traduction est une forme d'explication de la signification d'un texte. Les projets que j'ai eu à traiter au cours des dernières années portent sur la question de la traduction. Je prends donc en considération des logiciels qui, insérés dans notre application Web pour la linguistique computationnelle et la philologie, effectuent deux activités complémentaires :

1 Ce *modus operandi* est analogue à la théorie exprimée par Wellek et Warren en 1949 (Theory of Literature / *Théorie de la littérature*), selon laquelle un texte (littéraire) n'a pas une signification intrinsèque, ayant une nature "ouverte" à l'interprétation que chaque lecteur donne selon son expérience, ses connaissances, sa sensibilité, etc. L'objet textuel numérique, pas seulement littéraire, participe à une plus grande mesure de ce "relativisme" herméneutique.

- ils utilisent un texte et une traduction déjà disponibles. Les deux textes doivent être alignés pour faciliter les études historique, philologique et lexicale ;
- ils utilisent un texte de base et aident une équipe de spécialistes qui travaillent sur le Web afin de traduire l'original (dans ce cas, c'est le Talmud de Babylone en hébreu / araméen et la tradution en italien contemporain).

[Slides 28, 29]

Les diapositives 28 et 29 présentent un exemple du premier cas. Il s'agissait de mettre en comparaison certains chapitres des *Ennéades* de Plotin avec leur traduction en arabe, qui a été réalisée au IX^e^ siècle et qui est connue sous le nom de *Théologie* d'Aristote. Les deux textes ont été répartis en péricopes parallèles par des spécialistes dans le domaine des études historiques et philosophiques [**Slide 28]**. Les annotations les plus importantes sont relatives à l'évaluation comparative entre les termes arabes et les termes correspondants dans le texte original grec [**Slide 29]** : l'application doit permettre aux spécialistes de classifier ces annotations afin d'interroger la base de données (qui est constituée par des péricopes parallèles), non seulement par formes fléchies ou par lemmes, mais aussi par typologie d'annotation.

Le modèle avait été prévu pour gérer les annotations faites par des linguistes sur l'édition numérique des manuscrits inédits de Ferdinand de Saussure. Ce modèle a démontré sa robustesse et s'est adaptée à cette autre situation.

En conséquence, nous croyons que ce module peut faire partie d'une infrastructure web générique, destinée à produire des éditions critiques numérique.

Ce modèle applicatif n'a pas seulement un intérêt pour la recherche sur les données, mais aussi facilite la formation de nouveaux chercheurs qui apprennent, grâce à cette méthodologie, une nouvelle façon de travailler sur des textes numériques, dans le but de produire des éditions critiques.

[Slide 30]

Power Point seconda parte (Parigi_16_01_2014_II)

3. Texte et édition numérique : le rôle de la philologie computationnelle

[Slide 1 : lista dei contenuti con evidenziato il terzo punto]

L'édition numérique est par nature mobile, indéterminée, jamais établie. Ce fait a augmenté la valeur des témoins. L'accent s'est déplacé vers les documents, on a fait l'éloge de la variante contre le texte critique établi par un philologue. L'édition a même été considérée comme un artifice, avec une perte substantielle d'autorité relativement à la critique textuelle Lachmannienne[2]. Il est intéressant de noter que cette attitude n'est pas la conséquence directe d'un outil technologique nouveau, mais dépend d'une vision méthodologique tout à fait indépendante. L'objet numérique n'est pas la cause, il a simplement rendu plus facile la mise en ligne et la communication d'images numériques de toute une tradition manuscrite directe et indirecte et, sur cette base, de permettre la liberté d'étude ou de travail éditorial à quiconque. Il a déjà été soutenu que, en conduisant cette attitude à l'extrême, on démolit la critique du texte et l'essence de la philologie : on ignore la valeur d'un texte établi, au profit des différents témoins qui le transmettent.

À l'Institut de Linguistique Computationnelle du CNR de Pise, un projet a été lancé qui, tout en reconnaissant les grandes opportunités offertes par le développement du numérique, n'est pas destiné à délégitimer le travail d'analyse détaillée et de comparaison des sources afin de produire des éditions critiques dotées d'apparats de variantes.

[Slide 2 : elenco delle caratteristiche principali di COPhi]

Au cours de ces dernières années, en effet, a été étudié le modèle pour le développement d'une application philologique, appelée COPHI, acronyme de « Computational and Collaborative Philology », destinée à faciliter le processus de production des éditions critiques, diplomatiques ou interprétatives et constituée par plusieurs modules logiciels qui interagissent dans une architecture à composants (modularité).

Le modèle a été conçu pour que la même application puisse assister plusieurs chercheurs qui travaillent de manière collaborative (partage).

Une troisième caractéristique concerne la pluralité d'utilisations : en effet, nous voulons qu'une application Web moderne et efficace soit en mesure d'aider ceux qui travaillent sur des textes transmis par plusieurs sources, sur documents uniques, sur texte imprimé et enfin, sur manuscrits d'auteurs modernes et contemporains (flexibilité).

Le programme est ambitieux, mais le projet et, surtout, les outils de développement actuels, accompagnés de balisage normalisés, en permettent la réalisation (normalisation).

Afin que tout utilisateur puisse vérifier par lui-même les résultats, même s'il n'est pas membre du projet, aucun outil sous droit d'auteur n'a été utilisé ; tous les modules logiciels sont « *open source* ».

En résumé, le modèle de développement qui sous-tend le projet est fondé sur cinq principes :

2 Cf. Cerquiglini, 1989.

- modularité ;
- fonctionnement collaboratif ;
- flexibilité ;
- utilisation de balisages et de langages standards ;
- outils de développement en « open source ».

Grâce à ces principes et à l'évolution constante des technologies, nous projetons la mise en place d'une batterie d'outils de traitement, pouvant être définie comme une infrastructure technologique pour les « humanités numériques »[3].

[Slide 3 : lista dei sottosistemi di COPhi]

Les principaux composants du système de philologie computationnelle collaborative sont au nombre de six :

1. gestion et traitement des images numériques des sources ;
2. *preprocessing* et balisage du texte ;
3. contextualisation : aspects spécifiques de contextualisation manuelle ;
4. apparat critique ;
5. annotations non structurées, structurées et classées ontologiquement ;
6. *COPhi et la critique génétique*.

Ces sous-systèmes sont conçus de façon à pouvoir fonctionner indépendamment les uns des autres ou coopérer pour assurer que les actions de l'un d'eux, si nécessaire, ait un effet aussi sur les autres.

3.1 Traitement des images et traitement intégré texte/image

[Slide 4]

Le premier sous-système offre la possibilité de consulter le texte transmis par les sources, en feuilletant le catalogue des images à l'écran. Cette phase simule la lecture traditionnelle du texte effectuée directement sur les originaux, ou sur les reproductions photographiques (estampes, microfilms, CD). COPHI, cependant, peut fonctionner même en l'absence d'images numériques, c'est-à-dire qu'il a la capacité de traiter les archives textuelles qui en sont dépourvues. Il est essentiel que cette infrastructure de recherche philologique soit organisée sous une forme modulaire dans laquelle le module graphique de manipulation des images soit seulement l'une des fonctions disponibles. Si l'archive de données ne contient pas d'images, ou si l'utilisateur n'a pas jugé opportun de l'utiliser, ce module restera inactif, sans toutefois empêcher le fonctionnement des autres modules. Si, toutefois, les images sont disponibles et ont été chargées dans l'application, alors seront mises à disposition des fonctions spécifiques telles que

3 Pour obtenir une application flexible et réutilisable par un large éventail d'études de caractère philologique, l'architecture générale est basée sur le bien connu *Model-View-Controller* (MVC), qui sépare la représentation des données de la manière dont elles sont présentées ("*rendering*") et traitées ("*management*"). Voir, par exemple, le manuel technique Pitt 2012.

l'agrandissement, la variation du contraste, de luminosité, les niveaux de contraste ou de couleur, etc. La lecture du texte est ainsi facilitée.

La consultation d'images numériques pour l'établissement d'une édition critique apporte des possibilités d'analyse supplémentaires à la consultation des originaux ou des copies analogiques, par exemple la possibilité de tracer automatiquement des zones, allant même jusqu'à la segmentation automatique des parties de l'image correspondant à un mot.

[Slides 5, 6, 7, 8]

Un logiciel expérimental "*off line*" avec cette fonction a été développé il y a quelques années ; ce logiciel, en analysant l'histogramme avec la répartition des points-image (pixels), identifie les marges, les colonnes et les lignes d'écriture, et les parties internes aux lignes où s'inscrivent chaque mot. Grâce à cette procédure, le texte que le spécialiste transcrit est lié à la zone de l'image correspondant à un mot. Cette procédure, segmente virtuellement l'image numérique du feuillet manuscrit en une mosaïque dont chaque tuile est un mot.

[Slide 9]

Une concordance graphique (texte/image) a été réalisée afin d'évaluer si elle peut rendre plus simple et plus sûre la détection de fausses lectures (ou de variantes graphiques), en fournissant au philologue la liste de mots attestés une seule fois.

En consultant des contextes et des images, il a des preuves concrètes de comparaison en vertu desquels choisir des lectures alternatives, ou confirmer celles déjà effectuées. Ce logiciel expérimental « *off-line* » n'a pas encore été transformé en un composant de COPHI parce que le traitement de l'image des sources manuscrites anciennes, nécessaire pour obtenir les résultats présentés, exige des efforts techniques et un temps de réalisation très élevés.[4] [**Slide 10, 11]**.

Les possibilités de ces technologies ont évidemment des limites, par exemple lorsque le manuscrit est écrit en *scriptio continua* ou bien si le support matériel a subi de graves dommages. Le système n'est alors pas en mesure de détecter les valeurs de lumière et d'ombre correspondant aux bords de chaque mots, et par conséquent, à écrire les coordonnées spatiales dans une base de données. Dans ces cas là, les résultats obtenus ont été faibles et nécessiteraient l'utilisation de techniques plus sophistiquées (calcul parallèle, systèmes neuronaux, etc.).

[Slide 12, 13, 14 : segmentazione manuale delle immagini (papiro greco e ostraka demotici)]

Par conséquent, il a paru opportun d'étudier et de réaliser une solution alternative, moins coûteuse en ressources, mais cependant très utile : mettre à disposition un système de sélection manuelle des zones de l'image, de longueur et de taille variables en fonction du besoin spécifique, auxquelles peuvent être associés des annotations et des commentaires comme, par exemple, la description des raisons qui rendent incertaine la lecture et l'interprétation du texte manuscrit. Ces notes seront utiles et facilement accessibles surtout lorsqu'il faudra prendre des décisions pour établir le texte et enregistrer les informations dans l'apparat critique.[5]

4 Le cas des textes imprimés anciens est différent : le programme de segmentation automatique des images des différentes éditions des XV^{e}-XVI^{e} siècles du *Contradicentium medicorum* de Gerolamo Cardano a donné, par contre, d'excellents résultats (cf. Baldi, 2006), [cf Slide 8].

5 Un exemple qui confirme la validité de la relation entre texte et image dans un système informatique pour les études du texte est donnée par Corradini, 2007.

Le même outil de sélection qui permet de mettre en évidence une zone de l'image où il y a un mot, s'applique aussi si on a l'intention d'opérer sur des pièces plus larges tels que, par exemple, une section entière du texte avec des caractéristiques de forme ou de contenu dignes d'être mises en évidence ou dignes de mention (une deuxième main, une probable interpolation, etc.). Cette fonction du module (la sélection des zones sur l'image numérique), en conformité avec la modularité qui sous-tend l'ensemble du système, peut également être utilisée pour toutes les autres images des documents collationnés, dans le cas où, bien sûr, ils sont disponibles en format numérique.

Cette fonction peut aussi être activée selon le même principe, pour mettre en évidence des parties du texte transcrit, d'une part pour la segmentation manuelle de contextes (voir ci-après *la contextualisation manuelle*) ; d'autre part pour l'annotation, par exemple des termes qui dans un contexte donné présentent des valeurs sémantiques particulières, qui appartiennent à des langages spécialisés, qui sont les hapax, ou d'autres phénomènes similaires.

3.2 Pré-processing, marquage et traitement du texte transcrit

Le standard. En premier lieu, COPHI adopte le langage XML et, dans le cas spécifique du marquage du texte et des attributs qui lui sont associés, la version XML de la "Text Encoding Initiative (TEI)"[6]. Ce choix est désormais presque obligatoire, car la valeur des balises utilisées est univoque, et connue, parce qu'elle est décrite dans des recommandations facilement accessibles en ligne.[7] Les règles décrites ne sont certainement pas exhaustives, mais il faut décider à l'avance quels éléments du texte méritent d'être encodés afin de ne pas surcharger le travail préparatoire pour placer le document numérique sur le réseau. Il s'agit d'établir un juste compromis entre les résultats attendus (et, par conséquent, les éléments mineurs ou majeurs du texte à encoder) et les ressources investies en termes de temps et d'argent.

Par exemple : l'ordinateur sera en mesure d'effectuer l'indexation des mots d'une citation seulement si un codage approprié, également présent dans les recommandations de la TEI, est inséré au début et à la fin de la citation elle-même.[8]

Un point qui est considéré comme inévitable dans le développement de notre système est de mettre à disposition de l'utilisateur final une interface qui facilite le choix des balises et des opérations de codage. Un éditeur occupé par des problèmes scientifiques parfois très complexes ne

6 Un standard international dans le domaine du traitement du texte et des archives textuelles numérisées

7 Le manuel complet peut être téléchargé gratuitement à l'adresse: http://www.tei-c.org/Guidelines.

8 La question de la reconnaissance des phénomènes intertextuels ou des références expresses ou implicites est très actuel et, référé en particulier aux textes de l'époque moderne, coïncide avec celle du plagiat. Du point de vue informatique sont de plus en plus nombreux les logiciels qui analysent ce qu'on appelle " text re-use" grâce à des algorithmes statistiques qui analysent les contextes identifiés au sein d'une œuvre et les mettent en comparaison avec un large corpus textuel afin de mettre en évidence les similitudes qui dépassent un certain seuil. L'ordinateur est en mesure de présenter au philologue les passages qui pourraient avoir un lien direct avec les passages similaires dans la forme et le contenu, et qui, sans arriver à les considérer comme des preuves de plagiat, aident par contre à identifier les phénomènes d'intertextualité. Dans ces cas, le système computationnel introduit dans le texte, sans aucune intervention humaine, les signes de balisage qui délimitent les parties impliquées dans le phénomène et va leur associer l'information sur la source auquelle les parties semblent avoir des forts éléments en commun non seulement en termes de *signifiant*, mais, surtout, de *signifié*. À cette technique, on préfére plutôt une intervention directe du specialiste dans le cas où une citation est présente sous une forme explicite, dans laquelle on indique le nom de l'auteur et éventuellement le titre de l'ouvrage cité.

doit pas être distrait par la consultation d'un catalogue de balises pour choisir celle qui convient à chaque situation spécifique.

Tout en réaffirmant que la machine pourra fournir des résultats significatifs, que si et seulement si on y entre une information bien structurée ; il est aussi essentiel que le spécialiste travaille avec une méthode simple, et intuitive, qui l'aide à formaliser son questionnement avec des balises que la machine peut compter et analyser.

Il y a ainsi des centres très actifs dans le domaine de la numérisation de textes et de la production de systèmes de traitement linguistique qui, pour les raisons ci-dessus, ne suivent pas les recommandations de la TEI, mais utilisent des systèmes de codage plus simples et plus orientés vers l'utilisateur final. Le respect du standard, peut cependant être assuré, car à tout moment on peut créer des tables de correspondance entre les valeurs exprimées avec la codification simplifiée et celles imposées par la TEI.

Ce qui importe, toutefois, c'est le respect d'un critère qui, une fois établi, sera suivi rigoureusement et sans ambiguïté, de sorte que les données puissent être intégralement interprétées et même réutilisées dans le cadre d'une communauté d'utilisateurs qui, en nombre croissant, a décidée d'adhérer au standard TEI.

COPHI utilise 3 classes d'éléments distinguables du point de vue philologique : les éléments extra-textuels, les éléments para-textuels et le texte lui-même.

Les éléments de balisage qui permettent de distinguer les unités de manière appropriée en fonction de la classe à laquelle elles appartiennent permettent au logiciel de traitement de fournir des résultats sélectionnés.

Il s'ensuit, par exemple, que l'on pourra avoir un index alphabétique spécifique avec seulement les termes utilisés dans les citations, comme vous pouvez le voir dans les diapositives **15, 16 et 17**, en se référant au texte des Anciens Grammairiens Latins.

Nous croyons qu'il est arbitraire d'imposer une liste d'éléments pour chaque classe, car elle sera subjective : un élément que nous considérons comme de type para-textuel pour les caractéristiques de la source, du texte qu'elle contient et de nos spécifiques objectifs d'étude, pour d'autres chercheurs pourrait plutôt être évalué comme un élément extra-textuel, en raison des différentes perspectives de recherche. Nous sommes donc en train d'établir le travail de développement des éléments de *preprocessing* en rendant disponibles pour les utilisateurs seulement 3 conteneurs correspondant aux 3 classes indiquées (extra-textuel, para-textuel et textuel), en leur laissant la tâche et la responsabilité de remplir chacun d'eux avec les éléments souhaités. Le chercheur doit être conscient que plus il aura de balises dans chacune des 3 classes, plus il aura d'informations à interroger avec la machine.[9]

9 La liberté de choix et d'action que COPHI entend laisser au philologue nous a conduit à ne pas imposer le système d'encodage TEI qui peut impliquer une importante charge de travail supplémentaire. Nous avons préféré fournir une liste générale des situations exprimée en langage naturel et sous forme d'un catalogue où il ne sélectionera que ce qu'il jugera utile d'adopter. À chaque élément de la liste correspond un codage COPHI auquel, à son tour, est associé le correspondant codage TEI pour garantir le respect de ce standard reconnu et la conséquente interopérabilité avec toutes les données qui l'adoptent. L'interface doit rendre très simple la sélection des éléments dans le catalogue, même pour ceux qui n'ont pas d'expérience dans l'utilisation des outils informatiques. Le prix à payer pour vouloir laisser ce choix au philologue pourra être une équivalence imparfaite entre certaines fonctions

3.3 La contextualisation manuelle

Une attention particulière a été accordée aux modalités de balisage des contextes que le système informatique propose à l'étudiant ou au chercheur, afin d'être plus précis que ce qui est produit automatiquement par l'utilisation de la ponctuation comme élément de délimitation.

Le système développé par l'ILC, en plus de la contextualisation automatique, gère aussi la définition manuelle de contextes. Cet aspect souvent négligé mérite pourtant une attention toute particulière. Il s'agira, par exemple, des projets ayant pour but l'étude lexicographique menée sur des textes anciens, ainsi que leurs traductions anciennes. Un contexte du texte source, par exemple une phrase, peut être interprété selon une segmentation différente dans le texte traduit. Un système approprié d'organisation et de comparaison paut ainsi aider à mettre en évidence d'éventuelles parties interpolées.[10]

Ce composant additionnel consiste en un système de balisage manuel des contextes qui d'un côté permet d'éviter le caractère arbitraire de la délimitation effectuée par un logiciel, de l'autre côté offre l'avantage de montrer les péricopes définies par ceux qui ont été capables d'en évaluer le sens.

L'application, une fois encore, diffère des pratiques, désormais très communes, de génération automatique de concordances contextuelles qui ne tient pas suffisamment compte de la nécessité fréquente de découper "manuellement" les contextes. Déléguer cette tâche à l'ordinateur peut empêcher une bonne compréhension du rôle qu'une expression linguistique joue là où elle est insérée. Cela est particulièrement vrai dans le cas des œuvres de sujet technique (médico-pharmaceutique, mathématique, astrologique/astronomique, etc.), copiées et/ou traduites dans un environnement très différent de celui dans lequel elles ont été conçues, puis transmises à travers les siècles. C'est pour ça que COPHI fournit une fonction grâce à laquelle la responsabilité de la délimitation des contextes parallèles et, par conséquent, de leur alignement, peut être entièrement contrôlés par le spécialiste.[11]

[Slide 18 e 19 : pericopi parallele per interpretazione semantico-ermeneutica]

Traduction du texte grec : « Est-ce que, dans l'âme, *l'indivisible et le divisible* sont au même endroit, comme s'ils étaient *mélangés* ? Ou bien l'indivisible appartient-il à l'âme autrement et sous un autre rapport que le divisible, et le divisible est-il à la suite de l'indivisible, et forment-ils deux parties différentes de l'âme, au sens où nous disons que la partie raisonnable est différente de la partie rationnelle ? »

Traduction du texte arabe : « Et maintenant nous voulons rappeler la cause qui fait si que des noms différents sont imposés à l'âme et lui est rattaché ce qui se rattache à la chose qui est

exprimées par le système de codage COPHI et leur équivalent TEI: nous croyons, toutefois, que ce résultat peut être atteint progressivement, en considérant maintenant prioritaire l'exigence d'autonomie et la facilité d'utilisation.

10 L'outil de contextualisation manuel de COPHI peut trouver une application aussi dans le domaine de la critique génétique. Voir ci-dessous le paragraphe *Module pour les avant-textes* de la section *Extension du modèle*. Dans ce document, les portions de texte découpées à l'aide d'une procédure manuelle, selon les principes sémantiques et interprétatifs, sont appelées péricopes, tandis que celles produites par un logiciel, selon des critères purement mécaniques (ponctuation, nombre de mots) sont définies contextes.

11 Pour un exemple d'un travail effectué à l'aide d'un système de génération automatique des indices, mais avec un tracé manuel des contextes, lire Corradini, 1982.

particularisée à cause de sa **subdivision** essentielle. Il faut donc savoir ceci : **l'âme est-elle particularisée, ou bien elle n'est pas particularisée ?** Et si elle est particularisée, est-elle particularisée par son essence, ou bien par accident ? Également, si elle n'est pas particularisée, c'est par son essence qu'elle n'est pas particularisée, ou bien par accident ? Nous disons que l'âme est particularisée par accident, et ceci parce que lorsqu'elle se trouve dans le corps, elle reçoit la particularisation par le biais de la particularisation du corps, comme le discours qui dit que **la partie raisonnante de l'âme est différente de sa partie irrationnelle** et sa partie désirante est différente de la partie colérique. »

Comme vous pouvez le voir, le contexte arabe est plus large que l'original grec parce que le traducteur a transféré le sens afin qu'il puisse être bien compris dans la culture arabe. Dans des cas très complexes comme ceux-ci, il est impensable de déléguer à un logiciel la tâche d'aligner les deux textes.

3.4. L'apparat critique

Si la lecture des sources implique l'identification des variantes par rapport au témoin principal, le système ouvre une zone de l'environnement de travail dans laquelle il y a autant de champs-variants que de témoins collationnés.

[Slide 20, 21, 22]

La combinaison de ces champs prend l'aspect d'un système complet dans lequel apparaît un champ où l'éditeur peut écrire ses propres choix. Il dispose ainsi d'un outil nécessaire à la composition d'un apparat positif où il peut inscrire toutes les leçons des témoins, y compris celles adoptées par lui dans le texte critique. Par conséquent, l'application qui aide l'éditeur lors de la production d'une édition critique simule la façon traditionnelle de procéder.

L'éditeur est évidemment libre d'ignorer les leçons banales, non substantielles (par exemple, les variantes orthographiques), inutiles à l'élucidation des relations entre les témoins qui, par conséquent, ne seront pas incluses dans l'apparat. Néanmoins, il est conseillé, au moins pendant les premières étapes du travail, d'enregistrer fidèlement n'importe quel type de divergence, que ce soit les erreurs qui pourraient aider à déterminer le *stemma codicum* (comme, par exemple, les erreurs de transcription tels que l'aplographie, l'homoteleute, la dittographie, etc.) ou les leçons qu'on jugera ensuite opportun d'éliminer, comme par exemple les erreurs serviles des copistes, inutiles aussi comme indication des *scriptae* particulières. Afin de faciliter l'éditeur dans cette opération, l'interface de l'environnement de travail fournit une section pour enrichir chaque variante de l'apparat avec toutes les annotations.

Un bon système d'indexation de ces annotations permettra, par exemple, de trouver facilement toutes les leçons considérées comme des erreurs banales, de les évaluer globalement et de les éliminer complètement avec une plus grande conscience (voir point suivant).

Puisque chaque leçon insérée dans l'apparat peut avoir été déjà analysée dans des éditions antérieures ou dans les études critiques spécifiques, il est prévu de pouvoir insérer des informations bibliographiques ou une adresse web.

Comme nous l'avons souligné à plusieurs reprises, le système propose toujours un moteur d'indexation qui s'active, sur indication de l'éditeur critique, sur les sections qui au fur et à mesure

sont l'objet de l'étude et de l'évaluation. Il va produire, par conséquent, les index alphabétiques des leçons du texte critique, des variantes de chaque témoin et il est capable de connecter les uns (les leçons du texte critique) avec les autres (les variantes) et vice versa.

La création d'un apparat positif offre un certain nombre d'informations qui peuvent être traitées par le système informatique au bénéfice de l'éditeur critique qui s'est consacré à effectuer la transcription uniquement du témoin considéré comme le plus fiable (selon des évaluations internes et externes). Contrairement aux autres programmes informatiques d'édition critique, ce logiciel, au lieu de forcer l'utilisateur à produire la transcription de toutes les sources dignes d'être collationnées, exige de transcrire un seul témoin, mais est capable de générer automatiquement le texte transmis par les autres. En utilisant les parties du texte transcrit auxquelles ne sont pas associés de variantes et en les intégrant, petit à petit, avec celles qui sont attestées dans d'autres sources et que le logiciel retrouve dans les champs de l'apparat, il peut reconstituer le texte de tous les témoins. Si dans certains cas une telle procédure ne semble pas pleinement justifiée[12], dans d'autres cas on pourrait en revanche offrir des conditions de lecture séquentielle des sources considérées a priori comme pas ou peu fiables : en les regardant à nouveau dans leur intégralité, on pourra peut-être élucider les raisons sur lesquelles les doutes étaient fondés. Il est évident que si on fait fonctionner le moteur de recherche sur le texte de chaque témoin généré automatiquement, on peut mieux mettre en valeur les caractéristiques graphiques et phonétiques, ou les « espions de la langue » moins évidents à détecter de façon isolée dans l'apparat critique.

Ce *modus operandi*, en plus, répond à un nouveau besoin, né à une époque où s'affirme la culture numérique : le texte prend une dimension de plus en plus mobile, sans la rigidité que le papier impose, au moins pour un temps, jusqu'à ce qu'il soit réédité et réimprimé. Le fait qu'il puisse être généré par un programme informatique ne change pas, fondamentalement, les activités de ceux qui se sont engagés à étudier son histoire, sa tradition et à proposer une version vraisemblablement identique ou proche de celle écrite par l'auteur et qui, dans sa version originale, a été perdue. Cependant, ce qui change de plus en plus est la possibilité d'utiliser la qualité "numérique" du texte pour le soumettre à une analyse automatisée utile pour la recherche philologique. Le logiciel et ses produits ont donc tendance à réduire l'incertitude d'une décision (le *divinatio* de la méthode de Lachmann) au profit de données plus objectives.

On pourrait objecter qu'un philologue expert n'a pas besoin des données consistantes du point de vue numérique pour ses décisions : il lui suffit de peu d'éléments fondamentaux pour opérer une distinction entre les sources à collationner et celles à rejeter ou à examiner seulement partiellement. Qu'il ait imaginé ou non le *stemma*, les outils d'évaluation que lui offre un ordinateur, surtout si cela n'entraîne pas une charge de travail et de temps, sont très utiles. La génération du texte de chaque témoin à partir des informations de l'apparat est activable, bien sûr, aussi pour associer les choix éditoriaux aux leçons des autres témoins sélectionnés par l'éditeur : le résultat est le texte établi qui pourra être placé sur un serveur connecté au réseau, ou être imprimé.

12 Il s'agit de phénomènes tels que, par exemple, la présence de sources largement contaminées qui conduisent parfois à de traditions textuelles très complexes. Il en résulte une grande difficulté à reconstruire dynamiquement, en utilisant les fonctions du logiciel décrites ci-dessus, le texte de tous les témoins collationnés sur la base d'un texte unique.

3.5. Insertion d'annotations

3.5.1. Annotations non structurées

Cette fonction permet de sélectionner une partie du texte ou de l'image afin d'y associer une annotation. Les informations sont enregistrées dans une base de données et sont précieuses surtout pour d'autres chercheurs dans le cas où le travail est développé dans un esprit collaboratif. Il est bon toutefois d'insister sur le fait que :

[Slide 23] les annotations dont on parle ici sont exprimées sous forme non pas structurée, mais discursive ; elles prennent la forme de courtes études monographiques sur une expression, un mot, une variante, une utilisation particulière et inhabituelle d'un mot par rapport à l'utilisation communément connue. Le texte de ces brèves observations joue principalement le rôle d'une note provisoire qui n'est pas nécessairement destinée à être publiée en ligne ou dans sa version imprimée. Si le système est utilisé par une communauté de chercheurs pour un travail collaboratif, ces annotations servent à partager les remarques et les problèmes rencontrés par chacun afin que d'autres puissent en bénéficier.

3.5.2. Classification des annotations

COPHI a fourni aussi une possibilité supplémentaire de naviguer dans les données textuelles, en proposant un modèle de classification des annotations. Offrir au chercheur la possibilité d'organiser ses propres commentaires et annotations selon une typologie représentée par mots-clés (une sorte de catalogue par sujets) présente un avantage significatif, en particulier pendant la phase d'*information retrieval*. Grâce à cet outil qui est bien sûr facultatif, il est facile de demander au système de montrer toutes les parties du texte qui, sur la base du même type d'annotations attribuées par le chercheur, ont évidemment des traits en commun (par exemple toutes les leçons d'une source qui sont considérées comme des erreurs triviales, ou tous les passages qui contiennent des caractéristiques graphiques/phonétiques particulièrement importantes, etc.). La réalisation de cette procédure n'est pas techniquement complexe et n'implique pas, pour l'utilisateur une charge de travail et de temps : en fait, il sera en mesure d'entrer, dans un menu déroulant, un ensemble de mots clés ou de sujets qui seront utilisés pour classifier l'annotation qu'il jugera approprié d'introduire pendant le travail. L'opération d'insertion de mots clés est continue : la liste peut donc être mise à jour à tout moment, peut être partagée éventuellement avec d'autres chercheurs qui collaborent au même projet éditorial, ou au contraire, reste à l'usage exclusif du chercheur qui l'a composée.

Il est donc évident que, même pour cet aspect, on a été délibérément évité d'imposer une classification établie a priori, en préférant au contraire laisser à l'utilisateur le choix d'un nombre illimité de termes à utiliser comme mots-clés.

[Slide 24]

Le résultat produit par le système pour répondre à une requête spécifique (par exemple : montrer toutes les parties annotées comme "*misunderstanting*" = mésentente) permettra d'avoir une vue synoptique de tous les passages (contextes) qui ont été annotés avec cette typologie. Le

chercheur dispose de données plus nombreuses et mieux sélectionnées pour confirmer ou infirmer ses observations.

En d'autres termes et pour résumer ce qui a été dit à cet égard, l'application fournit deux formes différentes pour les annotations : une libre, et l'autre dirigée.

L'annotation libre implique la sélection d'un segment du texte (du mot à la phrase) et l'ouverture d'une zone de l'interface dans laquelle vont s'insérer les considérations dans une forme non structurée et sans contrainte d'espace.

L'annotation dirigée cependant, offre au spécialiste la possibilité d'énumérer une typologie de classes de jugement qui, incluses dans un menu, lui seront proposées chaque fois qu'il voudra marquer une expression particulièrement importante et digne d'être associée et comparativement évaluée avec toutes celles qui ont été classées de la même façon. COPHI ne fournit pas seulement une visualisation des correspondances sur les images (concordance texte/image), mais présente automatiquement tous les contextes annotés de la même manière.

3.5.3. Classification ontologique des annotations

Dans les deux sections précédentes, nous avons décrit deux façons différentes d'insérer des commentaires et des notes d'apparat critique. Une troisième méthode a été prévue pour permettre à un spécialiste d'organiser sémantiquement les contenus du texte qui méritent selon lui d'être mis en évidence, soit parce qu'ils sont intrinsèquement significatifs, soit parce qu'ils sont essentiels à sa recherche. Il s'agit, par conséquent, de fournir un système capable de représenter, sous une forme "ontologiquement" structurée, la connaissance d'un ou plusieurs domaines logiques et sémantiques que le texte critique établi (ou en cours d'édition) transmet. En bref, si les autres modules d'annotation sont principalement destinés à clarifier les décisions prises par le chercheur lors de la sélection des leçons pour la *constitutio textus* et pour la construction d'un apparat de variantes, cette troisième méthode vise plutôt à organiser les composants conceptuels que le spécialiste estime utile de mettre en évidence avec une attention particulière. La conception de ce troisième système innovant est illustré par un projet de recherche portant sur l'étude des éléments médico-pharmaceutique et anatomiques présents dans des manuscrits médiévaux écrits en occitan.

Ce qui a été dit précédemment à propos de la possibilité d'administrer une liste de mots-clés capables de classifier thématiquement les annotations, peut être évoqué ici parce que, dans le cadre d'une organisation sémantique plus précise des données, nous pourrons insérer chaque entrée de cette liste dans un schéma où apparaissent les entrées de plus haut niveau, les entrées sous-jacentes ainsi que les relations entre ces entrées et sous-entrées.

La connaissance du domaine spécifique que les textes transmettent est ainsi représenté sous une forme ontologiquement structurée et, par conséquent, explicite. L'évaluation précise de la terminologie médicale et pharmaceutique médiévale dans l'ancien occitan est une étude de cas particulièrement intéressante, car elle nécessite des outils d'analyse plus fins que ceux offerts par des simples indices de mots-formes ou de lemmes présents dans les sources, accompagnés des éventuelles concordances.

[Slide 25]

Il semble aujourd'hui de plus en plus fonctionnel d'interroger une base de données terminologique ou de textes en utilisant comme clé d'accès, un concept ou un thème générique (par exemple, "pommade"), ou une relation entre deux concepts (par exemple "pommade" pour

"blessure"), ou même une relation entre plusieurs éléments (par exemple,"pommade" pour "blessure" dans une certaine partie du corps, telle que la "tête").

Les résultats obtenus avec des programmes traditionnels pour produire les vocabulaires terminologiques des secteurs spécifiques pourraient d'ailleurs ne être pas exhaustifs, car on a constaté par exemple que :

- l'une des sources (ou quelqu'un qui effectue une recherche) désigne un même thème avec des termes différents de ceux utilisés dans une autre source ;
- dans la phase d'*information retrieval*, une clé d'accès (par exemple, "pommade", "blessure", "tête") sémantiquement équivalente mais non attestée, est utilisée. Ceci rend impossible la récupération d'informations qui sont pourtant bien présentes, mais sous une autre forme.

Le dépassement de ces limites et l'obtention de résultats complets sont rendus possibles par un système de "modélisation de données" (*data modeling*) qui aide l'utilisateur dans l'organisation et la rédaction du schéma conceptuel typique de son domaine particulier d'expertise et d'intérêt. La conception de ce système innovant a été réalisée pour le projet de l'édition électronique d'un corpus de manuscrits de F. de Saussure.

Donnons un exemple concret pour clarifier les raisons qui ont conduit à adopter cette méthodologie et à concevoir les composants de ce système informatique.

[Slide 26]

C'est le cas lorsque des termes différents subsument la même valeur sémantique insérée dans le schéma : par exemple, le concept de "pommade" dans l'ancienne occitan *oignement* (et ses variantes *oinhement, oinnement, onhement, hongemen*) et *onguent*, mais aussi *dura confeccio* et *apostolico*, qui sont des types spéciaux de pommade. Le problème pourrait être résolu en partie en utilisant d'autres méthodes telles que par exemple, la préparation de tableaux de correspondance permettant à l'ordinateur de comparer au moins les variantes (graphiques, phonétiques, morphologiques) d'un même mot. Toutefois, les caractéristiques linguistiques du corpus amènent la procédure à générer des erreurs : l'utilisation mécanique de concordances par un programme peut en effet faire considérer des variantes d'un même mot comme des formes qui en réalité appartiennent à des mots différents.

Avec l'attribution de la valeur conceptuelle "pommade" toutes ces difficultés sont surmontées et les deux lemmes *oignement* et *onguent* – avec tous les autres synonymes possibles et, bien sûr, toutes les variantes – sont unifiés sur une base sémantique et conceptuelle. L'ensemble des contextes dans lesquels apparaissent les lemmes peut être obtenu en choisissant l'entrée "pommade" dans le schéma ontologique établi.

Ces procédures empruntées au Web sémantique, ont un pouvoir expressif très intéressant et sont utiles surtout pour extraire des informations lexicographiques difficiles à trouver, que ce soit par le catalogage traditionnel ou par des concordances contextualisées.

[Slide 27]

Les groupements de termes associés sur une base conceptuelle permettent :

- de donner une structure sémantique homogène et de définir explicitement la terminologie du domaine ; donc de faciliter sa compréhension ;

- de détecter même les plus petites nuances de sens ;
- d'interroger conceptuellement la base de données textuelles ;
- le schéma conceptuel est indépendant de la langue qui décrit les classes, sous-classes et les relations ;
- le schéma conceptuel est indépendant de la langue de l'archive textuelle.

Sur la base des résultats positifs obtenus, on juge que cette perspective technologique a une valeur significative dans la préparation de cette terminologie spécialisée présente dans les textes, et pas seulement dans les textes anciens.

3.6. COPhi et la critique génétique

La nécessité de fournir un système adapté à une vaste gamme d'utilisations a conduit à l'extension du modèle à d'autres domaines d'études du texte numérique : le premier concerne la philologie génétique, le deuxième la philologie du texte ancien imprimé. La méthode de consultation des images et des textes, le module pour l'apparat et les annotations, la possibilité de découper les contextes avec une procédure manuelle ou automatique, et enfin les possibilités d'interrogation offertes par les nombreux index produits répondent aussi aux besoins des philologues qui étudient les textes manuscrits sur lesquels l'auteur lui-même est intervenu à plusieurs reprises.

Ce même module logiciel COPHI permet en outre de faciliter le travail d'étude d'une œuvre imprimée, en comparant les différentes éditions qui ont été publiées au cours du temps : cette situation, au moins d'un point de vue structurel, est comparable à la *collatio* de la critique textuelle des témoins manuscrits. Les outils que l'application fournit, par conséquent, peuvent également être appliqués à la philologie du texte imprimé numérique.[13]

C'est précisément pourquoi nous ne décrivons maintenant que la première des deux situations, en soulignant toutefois qu'il s'agit d'une hypothèse théorique de travail. Nous ne sommes pas encore en mesure d'indiquer, en effet, les limites d'applicabilité effective de cette composante du système à des situations parfois très complexes, où les avant-textes[14] se chevauchent, de façon souvent indéchiffrable et sans pouvoir comprendre la séquence temporelle qui les a déterminés.[15] Nous avons jugé utile, toutefois, de décrire les fonctions que COPHI pourra mettre à disposition dans ce domaine, car les résultats pourraient fournir une aide concrète à la connaissance du chemin tourmenté que l'oeuvre a suivi, avant d'arriver à la version imprimée autorisée par l'auteur. Paradoxalement, le modèle du système de philologie computationnelle réunit deux situations diamétralement opposées : d'un côté les phases (sous forme de copies réalisées par personnes différents, dans des lieux et à des moments différents) qui se sont succédées à partir d'un texte original qui a été perdu ; de l'autre côté les phases (sous forme d'avant-textes) effectuées par l'auteur lui-même dans des périodes antérieures à la version du texte que nous connaissons dans sa dernière version, parfois définitive. Comme nous l'avons dit, malgré le retournement de perspective, la structure logique du processus est très similaire et COPHI entend

13 Cf. encore Baldi, 2006.

14 Le terme « avant-texte » est ici utilisé dans la valeur attribuée par Segre, 1985.

15 Sur une évaluation précise et récente des problèmes liés à la critique génétique, lire de Biasi, 2011.

s’appuyer sur cet élément de similarité pour étendre son champ d’application de la critique textuelle à la philologie génétique.[16]

Module pour les avant-textes.

L’étude du modèle et l’actuelle logique de programmation prévoient que l’application contribue à la production d’éditions numériques de manuscrits autographes sur lesquels l’auteur est intervenu à plusieurs reprises.

De façon plus générale, nous avons l’intention de vérifier si, grâce à la structure du module qui gère les annotations, il est possible d’organiser et de traiter de façon appropriée la pluralité des avant-textes qui apparaissent, peut-être en se chevauchant, dans la même page d’un document et qui, dans de nombreux cas, ont précédé la publication d’une oeuvre.

[Slide 28, 29, 30, 31, 32]

Nous connaissons depuis longtemps la difficulté de travailler avec les variantes d’auteur en utilisant des techniques et des langages hypertextuels-multimedials avec lesquels on peut obtenir des résultats satisfaisants, mais au prix d’un effort excessif du philologue dans la phase indispensable de codage des textes, quand il s’apprête à effectuer la transcription de ce qu’il voit, ou est en mesure de voir, dans l’image du manuscrit original ou numérisé. Cette opération est cependant indispensable pour les programmes de traitement des hypertextes, car ils doivent reproduire le texte, souvent avec des motivations douteuses, en préservant l’aspect graphique du texte autographe.[17]

On juge que l’organisation des données sous forme de péricopes parallèles, auxquelles on peut associer des annotations et des commentaires, est une modalité plus efficace et beaucoup moins exigeante.

Le module « computationnel » consiste à considérer les différentes “versions” d’un même texte toutes comme des avant-textes divisés en péricopes par le chercheur, ensuite analysés par le système. En résulte une table dans laquelle les colonnes représentent les phases successives du réarrangement, tandis que les cellules contiennent le texte des péricopes. Une portion de texte effacée par l’auteur apparaîtra dans deux cellules alignées et faisant partie de deux colonnes différentes : la première cellule contiendra le texte, tandis que la cellule correspondante sera vide. Au contraire, un ajout sera présenté sous forme de deux cellules adjacentes dont la première sera vide, alors que l’autre contiendra la partie du texte ajouté plus tard. On a autant de colonnes et, par conséquent, autant de cellules de texte que de phases de réécriture opérées par l’auteur.

---

16 Des phénomènes similaires se produisent dans les textes imprimés anciens comme, par exemple, *La Scienza Nuova* de Vico, dont on connaît 63 copies riches de notes manuscrites autographes. L'édition critique moderne (cf. Cristofolini, 2004), naturellement, enregistre dans l’apparat de telles annotations, mais il est évident que le support papier limite fortement les formes de consultation, que seul le support électronique peut assurer. Dans ce cas, la tâche principale est de faciliter la lecture parallèle entre le texte imprimé et le texte des notes manuscrites correspondantes qui ne sont pas toujours identiques dans les différents copies, et qui consistent tantôt en des interventions interlinéaires, tantôt en des vraies annotations de longueur variable. Certaines simulations ont montré largement qu'elles peuvent être traitées avec le même module de démarcation des péricopes et des annotations afférentes décrite ci-dessus.

17 Cf. par exemple, D’Iorio, 2000.

Dans certaines limites imposées par la lisibilité de ce qui a été effacé, on pourra effectuer des lectures séquentielles des différents avant-textes en fonction de la colonne utilisée par le système, sur demande du critique, pour joindre les péricopes contenues dans ses cellules.

En un certain sens, le système est capable de régénérer les avant-textes en agglutinant la séquence des péricopes et en facilitant l'étude du processus génétique qui a conduit à la création d'une œuvre littéraire. Le spécialiste en philologie génétique aurait a disposition des éléments exhaustifs et bien structurés pour examiner les raisons stylistiques, linguistiques, psychologiques qui ont conduit l'auteur à intervenir sur son œuvre avec des effacements, ajouts interlinéaires, ajouts marginaux, notes de bas de pages, etc.

En bref, le modèle prend en compte les variations de l'auteur, les avant-textes (structurellement assimilables à des traductions de parties d'un texte), et la comparaison entre ces traductions permet d'introduire des évaluations, des annotations critiques, sémantiques et interprétatives.

Conclusion

Ce que j'ai indiqué ne doit pas être considéré comme un système complet, définitif. Mon intention était de présenter une approche méthodologique pour réaliser un système complexe de philologie computationnelle des documents numériques, adapté au travail collaboratif via le réseau. Une partie de ce système a été réalisée principalement grâce au financement de la Commission européenne à partir du 3e programme-cadre. Les activités menées sur la base de ce financement ont essayé de réaliser ce qui avait été prévu dans le cadre de chaque projet, mais ont aussi guidé le développement d'une application capable de répondre à des besoins plus larges. En ce sens, le modèle étudié et suivi dans les différentes étapes du développement, est caractérisé par une grande flexibilité d'emploi. Même si elle n'a pas toutes les conditions requises par une infrastructure technologique pour les sciences du texte, l'application et l'ensemble des modules réalisés (et encore à réaliser) constituent un premier pas dans cette direction. Personnellement, je ne pense pas que les infrastructures de recherche existantes pour les Sciences Humaines (par exemple CLARIN, DARIAH, etc.) aient suffisamment pris en compte les exigences des critiques textuels et, plus généralement, des philologues. J'ai pour cela une grande confiance dans la collaboration entre les centres de recherche prêts à travailler au renouvellement de ce domaine des Sciences Humaines, que ce soit pour la recherche ou l'enseignement et la formation de nouvelles générations de chercheurs.