“ Quels corpus et quels outils d’exploitation de corpus pour les études de linguistique et philologie romanes : l’unité de la romanistique ”

Pierre Kunstmann, Andrea Bozzi, Giovanni Parodi, Jean-Marie Pierrel, Achim Stein

« Quels corpus et quels outils d'exploitation de corpus pour les études de linguistique et philologie romanes : l'unité de la romanistique »

Pierre Kunstmann, professeur émérite à l'Université d'Ottawa, Andrea Bozzi, directeur de l'Istituto di linguistica computazionale «Antonio Zampolli», Pise ; Giovanni Parodi, professeur à l'Université pontificale catholique de Valparaíso ; Jean-Marie Pierrel, professeur à l'Université de Lorraine ; Achim Stein, professeur à l'Université de Stuttgart.

1. Introduction (Pierre Kunstmann)

Quand les organisateurs de notre congrès m'ont demandé, l'an dernier, d'animer la présente table ronde sur les questions « Quels corpus et quels outils d'exploitation de corpus », il s'agissait bien entendu dans leur esprit de corpus et d'outils électroniques. Non que les corpus statiques imprimés soient aujourd'hui sans valeur : ils sont encore utilisés, quoique rarement. J'avoue d'ailleurs avoir commis moi-même un travail de ce genre. Petite anecdote : il y a plus d'une trentaine d'années, ma *Concordance analytique de la Mort le Roi Artu*, fut publiée aux Éditions de l'Université d'Ottawa en 2 gros volumes, soit un total de pas moins de 2000 pages. Le regretté Kurt Baldinger, l'un des phares de la linguistique romane, dans un compte rendu qu'il fit paraître de mes deux volumes dans la revue de notre société, parla de *heidenarbeit*, travail de bénédictin. L'expression ne manqua pas de me surprendre, car j'avais alors pris résolument le parti du numérique et je procédai avec une vitesse et une précision que notre confrère n'avait sans doute pas pu imaginer. Nous assistions alors aux nouvelles *Noces de Philologie et de Mercure*, à l' « alliance heureuse du génie informatique et de l'érudition textuelle »[1]. Les nouvelles technologies ont apporté au linguiste et au philologue des aides considérables et des perspectives nouvelles, qui peuvent dérouter ou donner le vertige. Si elles ne sauraient remplacer le jugement critique de l'expert, celui-ci, de son côté, ne saurait non plus se passer d'elles à chaque étape de sa recherche.

Aussi avons-nous choisi d'aborder cinq grandes questions qui nous semblent cruciales pour nos travaux, nos disciplines et nos champs d'investigation. Sans vouloir empiéter sur les contributions de nos intervenants, j'exprimerai brièvement et en termes généraux, pour chacun des thèmes retenus, ses caractères principaux et ses enjeux. Question préliminaire toutefois : qu'est-ce qu'un corpus ? Pour la plupart des spécialistes, ce n'est pas seulement un ensemble de mots ou de textes, mais un regroupement structuré suivant des critères linguistiques. Le corpus, à proprement parler, se distingue donc de la simple base textuelle ainsi que de l'hypertexte, dont le système le plus vaste est la toile d'araignée mondiale, le WWW. Pour ma part, j'adopterai la définition proposée par F. Rastier (2004, 32) : « Un corpus est un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages, et rassemblés : (i) de manière théorique réflexive en tenant compte des discours et des genres, et (ii) de manière pratique en vue d'une gamme d'applications. » Définition qui me permet, par une transition naturelle, d'introduire les deux premiers thèmes de la table ronde.

1.1. L'annotation de corpus

L'annotation de corpus : dans les textes ou autour des textes. Le balisage interne constitue un apport hautement appréciable ; mais il n'est pas indispensable : un corpus brut est parfois préférable à un corpus mal codifié. La première étape de la codification est la lemmatisation, opération jadis très longue quand elle était effectuée manuellement, mais beaucoup plus rapide de nos jours avec l'aide de l'ordinateur. Les résultats sont presque parfaits pour les textes contemporains, mais ils sont aussi, dans certains cas, remarquables pour les textes anciens. Je tiens à mentionner l'outil conçu et développé par G. Souvay, au laboratoire qui nous accueille : LGeRM (Lemmes, graphies et Règles morphologiques), qui permet de « faciliter la consultation d'un dictionnaire, l'interrogation et la lemmatisation de textes médiévaux et trouve des applications dans l'édition électronique de manuscrits et la construction automatique de glossaires. » La deuxième étape du balisage interne est l'étiquetage morphologique et parfois aussi syntaxique : le premier est relativement facile à effectuer, le second constitue une opération plus complexe. Enfin le balisage externe est absolument indispensable : protocoles de sélection et d'établissement de documents, constitution d'en-têtes permettant une identification précise des documents et de leurs auteurs. Qu'on me

[1] Kunstmann 2012, 51.

permette de citer, à titre d'exemple, le travail d'une collègue de mon université, F. Martineau : son corpus MCVF (Modéliser le changement : les voies du français) comprend deux millions et demi de mots couvrant les périodes de l'ancien français au français classique, annoté suivant le protocole TEI et étiqueté pour la morphologie et la syntaxe, avec en-tête permettant une identification des textes et des auteurs. Le corpus est accessible en ligne sur le site du projet (www.voies.uottawa.ca).

1.2. La comparaison entre corpus

Il importe de distinguer si elle s'effectue dans le cadre d'une même langue ou entre deux ou plusieurs langues romanes. Dans le premier cas, on s'intéressera surtout au phénomène de la variation : les quatre types traditionnels de Coseriu (diachronique, diatopique, diastratique, diaphrasique), auxquels s'ajoutent la différence entre écrit et oral, celle entre les différents canaux de transmission de la parole (les nouvelles technologies brouillant souvent la limite entre oral et écrit) et même la distinction entre les sexes (diagénique). En ce qui a trait à la comparaison entre deux ou plusieurs langues romanes, il serait bon de s'interroger sur l'utilité respective des corpus parallèles (le même texte dans plusieurs langues) et des corpus comparables (différents textes de plusieurs langues, qu'on peut rapprocher suivant plusieurs paramètres).

1.3. L'édition de textes et de documents sonores

L'édition de textes et de documents sonores pose des problèmes considérables. Après la publication récente des meilleurs *tweets* de B. Pivot (*Les tweets sont des chats* 2013), on ne saurait négliger le « gazouillis », ce type d'énoncé si répandu maintenant de Bucarest à Valparaiso. Comment aborder aussi la production textuelle sur le Web ?

1.4. La pérennisation des documents

La perpetuazione dei documenti è un problema cruciale ed anche angoscioso per coloro i quali non vogliono lavorare con materiali effimeri: come conservare, diffondere, transcodificare ciò che ha richiesto delle settimane, dei mesi, ed anche spesso degli anni di lavoro ?

1.5. Les études linguistiques portant sur les langues romanes

Por fin, ya que los organizadores quisieron que subrayáramos la unidad de la romanística, conviene examinar la importancia de los corpus multilingües para los estudios contrastivos en la área de la Romania (Europa, África y Américas) e interrogarse sobre el papel de nuestra Sociedad para una mejor coordinación de los estudios romances.

2. L'annotation de corpus

L'annotation de corpus : dans les textes (lemmatisation, étiquetage morphosyntaxique, gloses diverses) et autour des textes (protocoles de sélection et d'établissement des documents, constitution d'en-têtes permettant une identification précise des documents et de leurs auteurs). Vers une standardisation.

2.1. Point de vue d'Andrea Bozzi

Il tema dell'annotazione dei testi, in realtà, permette di affrontare un tema ancora più vasto, ovvero il rapporto fra la Linguistica Computazionale, che da anni è impegnata nello studio di procedure di analisi e annotazione, e gli studi di carattere filologico-computazionale, nella prospettiva di mettere a disposizione dei ricercatori un'infrastruttura adeguata alle proprie attività. L'elemento centrale riguarda l'adozione di standard per la codifica e la rappresentazione dei dati. Si tratta di un tema molto sentito da tempo e discusso anche a livello Europeo: lo *Standing Committee for the Humanities* della *European Science Foundation* (ESF), per fare solo un esempio, ha prodotto un documento interessante due anni fa, utile e ben documentato[2]. Il fatto di avere a disposizione sistemi di codifica standard, come la TEI (*Text Encoding Initiative*, oggi disponibile anche nel formato xml)[3], aumenta la possibilità di interscambio dei dati realizzati in centri diversi dal momento che l'analisi linguistica e filologica, ivi comprese le informazioni strutturate

[2] Si veda *ESF Science Policy Briefing 42*, *Research Infrastructures in the Digital Humanities*, scaricabile gratuitamente dal sito della ESF (www.esf.org) col nome di: spb42_RI_DigitalHumanities.pdf

[3] Il manuale completo è disponibile all'indirizzo: <www.tei-c.org/Guidelines>.

degli apparati critici e della bibliografia, segue criteri di marcatura espliciti e condivisi. E' tuttavia necessario sottolineare due aspetti della questione, in particolare per quanto riguarda la TEI:

– la molteplicità delle attività e delle finalità di studio e di ricerca sui testi rende molto improbabile ottenere una standardizzazione esaustiva e da tutti condivisa. E' illusorio pensare che gli interventi, nel campo delle varie discipline filologiche, siano suscettibili di una strutturazione preventiva costringendo, di fatto, i singoli ricercatori a seguire comportamenti e norme completamente definite a priori. Questo atteggiamento, purtroppo, è stato ed è ancora vivo nella visione universalistica dei cultori della TEI e, a mio giudizio, esso non favorisce la conversione al digitale di coloro i quali ancora indugiano a servirsi di mezzi di indagine tradizionali anche su documenti già disponibili in formato numerico grazie all'aggiornamento tecnologico che si è verificato nelle biblioteche e negli archivi di tutto il mondo.

– Gli attuali sistemi di marcatura (come, appunto, la TEI che è diffusa a livello planetario) sono poco orientati all'utente finale inesperto e necessitano di una serie di interventi per migliorare il modo con cui essi siano impiegati senza l'ausilio di personale esperto. E' impensabile che chi si trovi ad affrontare problemi linguistici o editoriali talvolta molto complessi distolga il proprio impegno per accedere a liste di marcatori fra i quali scegliere quelli prescritti per ogni specifica situazione. Pur ribadendo che la macchina potrà fornire risultati importanti solo se disporrà di parti codificate bene ed in forma omogenea, è tuttavia indispensabile che si metta lo studioso nelle condizioni migliori e meno defatiganti per intervenire in questa delicata fase di preparazione dei dati. Egli deve disporre di un menù a forma di catalogo per selezionare, con la sola pressione del tasto del mouse, la voce da utilizzare e deve essere compito del sistema inserire il codice di marcatura ad essa corrispondente nella posizione indicata della versione digitale del testo.

Una ulteriore interessante prospettiva che si offre ai ricercatori è costituita, come si diceva, dalle infrastrutture di ricerca dove, tuttavia, deve ancora essere maggiormente rappresentata la comunità di studiosi che operano nel campo della filologia e della critica del testo. Ad oggi, infatti, né CLARIN né DARIAH[4], le due infrastrutture più sviluppate nel campo delle Scienze Umane, sono in grado di fornire risposte sufficienti ai bisogni della filologia del testo, nonostante l'enorme sviluppo della tecnologia digitale imponga urgentemente di trovare soluzioni adeguate in proposito.

2.2. Point de vue de Giovanni Parodi

En el marco de esta mesa redonda, resulta muy oportuno reflexionar acerca de los requisitos que se debe contar para la constitución de un corpus. Así, un referente importante es la propuesta de EAGLES (1996a y b), en la que se propone recomendaciones para que un corpus pueda considerarse como tal: 1) El corpus debe ser lo más extenso posible de acuerdo con las tecnologías disponibles en cada época, 2) Debe incluir ejemplos de amplia gama de materiales en función de ser lo más representativo posible, 3) Debe existir una clasificación intermedia en los géneros entre el corpus en total y las muestras individuales, 4) Las muestras deben de ser tamaños similares, 5) El corpus, como un todo, debe tener una procedencia clara.

En esta misma línea, algunos autores sugieren cuatro ventajas para adoptar una aproximación basada en corpus: 1) Adecuada representación del discurso en su forma de ocurrencia natural en muestras amplias y representativas a partir de textos originales, 2) Procesamiento lingüístico (semi) automático de los textos mediante el uso de computadores (Biber, Reppen, Clark & Walter, 2001). Ello permite análisis más amplios y profundos de los textos mediante conjuntos de rasgos lingüísticos caracterizadores, 3) Mayor confiabilidad y certeza en los análisis cuantitativos de los rasgos lingüísticos en grandes muestras de textos, 4) Posibilidad de resultados acumulativos y replicables. Posteriores investigaciones pueden utilizar los mismos corpus u otros pueden ser analizados con las mismas herramientas computacionales.

Ahora bien, ya sea si el corpus debe ser necesariamente de tipo digital o si aun es factible pensar en un conjunto de textos en papel, otros rasgos resultan más relevantes. También se hace evidente que el asunto de la extensión cobra importancia. Seguramente se dirá que ello depende en gran medida de los objetivos de la investigación. Sin embargo, sin importar que rasgos se desea privilegiar, si se busca un proceso de investigación sinérgico con resultados de índole acumulativa y posibilidad de replicación, resulta indudable que se debe adherir a la mayoría de las indicaciones propuestas.

Desde este marco se pueden identificar en mi opinión, al menos, ocho características relevantes, llegado el momento de construir y comprender los alcances de un corpus. Ellas se listan a continuación sin mediar ningún sesgo jerárquico. Al mismo tiempo, cabe destacar que este conjunto no está cerrado: 1) Extensión, 2) Formato, 3)

[4] Si vedano i siti ufficiali dei due progetti <www.clarin.eu> e <www.dariah.eu>.

Representatividad, 4) Diversificación, 5) Marcado o etiquetado morfosintáctico, 6) Procedencia, 7) Tamaño de las muestras, y 8) Clasificación y adscripciones de tipo disciplinar, genérico, temático, etc. (Parodi, 2010a).

Estas características se constituyen en principios a tener en cuenta, dependiendo de los objetivos de cada investigador y de las posibilidades tecnológicas al alcance. No obstante lo anterior, el concepto de corpus muestra una característica que se hace relevante y pone en evidencia ciertas complejidades: aquella denominada representatividad. Es bien sabido que incluso los grandes corpus no logran dar cuenta de la lengua como un todo ni tampoco se pretende que así sea. La lengua en su dinamismo y heterogeneidad es mucho más rica de lo que se puede imaginar y no logra ser captada en un solo corpus, por gigantesco que sea su tamaño. Como se sabe, un corpus puede ofrecer información detallada acerca de una lengua particular, pero es imposible recolectar un corpus que abarque toda una lengua. Si ese fuera el caso, sería necesario recolectar todos los usos de dicha lengua. De este modo, se debe siempre tener presente que un corpus es sólo una colección finita de, en el caso de la lengua, un universo infinito.

Ahora bien, en el foco del tema central, desde el punto de vista de la anotación, sin duda, el contar con etiquetadores morfosintácticos automáticos o semi-automáticos se vuelve un asunto fundamental. Un tema central en este punto lo constituye el tipo de gramática que subyace a esos marcajes morfosintácticos y los sustentos tanto teóricos como empíricos que dan origen a las categorías y nomenclaturas empleadas. Como se sabe, no existe un criterio conceptual unívoco de gramática y esta posible diversidad da origen a modelos diversos, hecho que afecta los tipos de *taggers* y *parsers* existentes.

Paralelamente, existe un tipo estadístico computacional que se ha hecho muy popular en los últimos años. Este es conocido por su sigla en inglés: LSA (*Latent Semantic Analysis*). Este sistema de análisis semántico, que no presta atención a la morfología ni a la sintaxis, opera sobre grandes cantidades de textos planos no marcados y fue desarrollado inicialmente para el idioma inglés (Deerwester, Dumais, Furnas, Landauer & Harshman, 1990; Dumais, 1994). En la actualidad también se cuenta con desarrollos para el español (Venegas, 2005, 2007; Gutiérrez, 2005).

Como se aprecia, contar con marcadores morfosintácticos presenta una serie de interesantes desafíos, los cuales han ido encontrando lentamente soluciones. También es posible trabajar con corpus no-marcados y se consiguen interesantes análisis. El problema de qué gramática emplear para que soporte los etiquetadores permanece como un factor que debe superarse para que la estandarización de principios pueda lograrse. Del mismo, las características de qué es un corpus y cómo conformarlo también debería buscar principios de estandarización, con el fin de que los datos recabados sean efectivamente comparables.

2.3. Point de vue de Jean-Marie Pierrel

Le coût de définition et de production de ressources linguistiques de qualité (corpus, corpus annotés, dictionnaires et lexiques) est important et c'est un gâchis de vouloir, pour chaque projet de recherche, redéfinir l'ensemble des ressources dont on a besoin. À titre d'exemple, l'établissement d'une base de données textuelle telle FRANTEXT[5] (Bernard, Lecomte, Dendien et Pierrel 2002) s'est chiffré en dizaines de personnes-an alors même que les annotations sur cette base demeurent minimales. Il convient donc de prendre conscience que sans une véritable mutualisation de telles ressources, chaque équipe de recherche ou chaque chercheur se verrait dans l'obligation de tout réinventer, alors même que nul ne peut être spécialiste de tout. Ceci est particulièrement vrai dans un domaine aussi vaste que la linguistique ou la philologie des langues romanes qui nécessite d'aborder des aspects aussi divers que le lexique, la syntaxe, la sémantique, la pragmatique.

Une telle mutualisation de ressources implique tout à la fois le développement et l'usage de normes ou de standard pour les ressources linguistiques et textuelles, mais aussi une prise de conscience de cette nécessité d'une meilleure mutualisation des ressources (corpus, dictionnaires et lexiques).

2.3.1. Concernant ces aspects « normes et standards »

Aujourd'hui, dans le cadre des documents textuels numériques, XML s'impose, tant pour les métadonnées que pour les données. Mais au-delà d'une simple utilisation d'un langage commun, il convient aussi de partager, voire normaliser, les schémas de balisage (ou DTD). C'est en effet au niveau de ces schémas qu'on est en mesure de définir proprement les balises utilisées, leurs attributs et leurs valeurs. Sans recommandations précises sur des standards à utiliser, chaque équipe risque de définir ses propres balises et schémas et la mutualisation ou la réexploitation de ce balisage par une autre équipe risque donc d'être difficile, voire impossible. Deux recommandations s'imposent donc :

a) Expliciter le mieux possible et de façon très précise son schéma de balisage.

b) Si possible le définir en accord avec les recommandations internationales.

[5] www.atilf.fr/frantext

Un travail important est mené au niveau international dans le cadre du comité technique TC 37 de l'ISO (le sous-comité dédié aux ressources linguistiques et à leur normalisation (SC4 : http://www.tc37sc4.org)) ou dans le cadre de la Text Encoding Initiative (TEI : www.tei-c.org), consortium international qui définit des recommandations de codage de ressources textuelles. Notre laboratoire ATILF participe à ces deux initiatives.

Mais ne nous leurrons pas, nous ne sommes pas encore à un stade où ces recommandations répondent entièrement aux besoins très diversifiés des chercheurs. Très souvent nous sommes confrontés à la nécessité de définir des extensions à ces schémas ou recommandations. Il est alors nécessaire de coopérer au sein de ces consortiums internationaux pour tout à la fois faire évoluer ces recommandations et faire prendre en compte les spécificités de nos besoins.

2.3.2. Conditions pour un meilleur partage des ressources (corpus, lexiques et dictionnaires)

Mais une des premières conditions à un vrai partage et à une mutualisation de ressources (corpus annotés, lexiques, dictionnaires) réside aussi dans un changement de mentalité. Trop souvent encore nous rencontrons des chercheurs faisant preuve d'une frilosité quasi maladive pour partager leurs ressources linguistiques et corpus. Les justifications qu'ils nous donnent sont très diverses, mais toutes aussi peu convaincantes. On peut citer entre autres :

– La crainte d'être dépossédés de leurs productions. Beaucoup de chercheurs restent « assis » sur leurs productions, refusant de diffuser leurs résultats ou ne les diffusant que partiellement, rendant ainsi leur réexploitation pas d'autres quasi impossible.

– L'espoir, souvent vain, d'une valorisation contractuelle ou industrielle de leurs résultats. C'est en particulier le cas pour de vastes ressources lexicographiques ou lexicales et cela conduit à un gâchis énorme : souvent plusieurs équipes sont amenées à refaire un travail déjà fait par d'autres. On pourrait, à ce propos, se poser la question de savoir s'il ne conviendrait pas pour la langue d'opter pour une philosophie comparable à celle retenue en génomique qui conduisit à rendre accessible à tous le décodage du génome humain.

– La volonté d'une publication papier avant toute ouverture de ces ressources, en particulier pour les travaux d'édition de textes anciens. Si ce souhait est à première vue le plus défendable, il convient de noter que les clauses contractuelles d'édition chez des éditeurs privés empêchent souvent ensuite une valorisation et un partage informatique de ces ressources via le Web.

Une recommandation s'impose donc : définir des métadonnées précisant les divers contributeurs à ces ressources et respecter une éthique scientifique en ne s'appropriant en aucune manière les travaux d'autrui. Cela est possible à condition de veiller à :

– Enrichir sans jamais les dénaturer les métadonnées d'une ressource que l'on peut être amené à enrichir ;

– Citer systématiquement les ressources utilisées dans nos travaux de recherche. Pour ce faire, une identification précise de ces ressources au travers d'un identifiant persistant de données ou PID[6] est alors nécessaire.

Sous ces conditions on peut espérer à terme créer une adhésion véritable aux efforts de mutualisation de ressources linguistiques dans des entrepôts fiables de ces données.

2.4. Point de vue d'Achim Stein

La standardisation de la représentation textuelle proprement dite me semble être chose faite : les standards XML et TEI sont largement répandus, et la plupart des outils de traitement et d'exploitation récents utilisent ces formats. Les questions que j'aimerais soulever sont en partie techniques, mais surtout linguistiques, puisque je suis utilisateur, non pas développeur dans ce domaine.

Sur le plan technique, il est important de souligner que c'est justement le progrès dans le domaine de la standardisation qui permet d'ajouter à un même texte des annotations multiples et multicouches : transcription de l'oral (segmentale et suprasegmentale), standardisation d'un dialecte, lemmatisation, étiquetage morphologique, annotation syntaxique, balisage des « entités nommées », etc. Certes, ces possibilités soulèvent également des questions : techniques, comme celle des conflits de segmentation ou de hiérarchie entre les couches d'annotation ; linguistiques, comme celle de l'interdépendance ou de l'indépendance des couches d'annotation. Mais d'un point de vue linguistique, la conséquence la plus importante est sûrement la mise en place d'une pluralité, voir d'une « convivialité » : l'état de l'art permet à plusieurs modèles, et même à plusieurs interprétations linguistiques, de coexister au sein d'une même ressource.

[6] http://www.gbif.fr/pid/presentation/db.html

Sur le plan linguistique, j'aimerais évoquer la relation entre le standard d'annotation et le modèle linguistique. Tout d'abord, « système d'annotation » n'équivaut pas à « modèle linguistique », mais les affinités entre standards d'annotation et modèles linguistiques sont parfois fortes, par exemple dans l'annotation phonétique entre la transcription de l'intonation et le modèle TOBI, ou bien dans l'annotation syntaxique entre la grammaire à constituants et le modèle créé par la University of Pennsylvania (UPenn), ou encore entre les corpus arborés dépendanciels et le modèle CoNLL. Ces modèles sont largement répandus et ils ont généré un certain nombre d'outils de requête et de traitement, sans pour autant recourir au format XML. Dans certains cas, il n'est même pas possible de conserver les informations éventuellement encodées en XML ou TEI dans le texte brut.

Je constate donc que, dans le domaine des corpus, il se reproduit plus ou moins ce qui s'est passé dans le domaine du développement des logiciels d'application : nous sommes témoins de la naissance de certains standards « de fait » (plutôt que « de droit »), établis par des projets de recherche de plus ou moins grande envergure et maintenus par les outils qui, souvent, ont été développés parallèlement aux ressources textuelles.

3. La comparaison entre corpus

Dans le cadre d'une même langue selon diverses considérations (époques, régions, styles, situations... ; oral / écrit), pour mieux cerner la variation. Entre deux ou plusieurs langues romanes : utilité respective des corpus parallèles et des corpus comparables.

3.1. Point de vue de Giovanni Parodi

Con el fin de permitir una adecuada comparación entre corpus, se hace importante que los corpus y sus textos constitutivos se encuentren debidamente identificados y catalogados. Esto es, un corpus debe mostrar más de alguna clasificación de la colección que recoge, ya sea de índole temática, de registro, de género o de disciplina. Tanto para fines de investigación como de almacenamiento, todo corpus debería estar descrito y sus rasgos deberían estar disponibles y accesibles.

En este contexto, la constitución de un corpus debería, preferentemente, contar con la posibilidad de disponer de otros tipos de corpus de naturaleza diversa en alguna dimensión. Ello permite la comparación y, de este modo, el contraste hace emerger características distintivas y prototípicas que, de otro modo, sería imposible llegar a descubrir. En este sentido, la recolección de un solo y muy focalizado corpus, por amplio que sea, no brindará una gran riqueza en su descripción, salvo que ya se cuente con otros corpus disponibles previamente y, así, la comparación emerja con mayor facilidad. Esta perspectiva de estudio ha dado origen a lo que se ha llamado el enfoque multidimensional.

El Análisis Multirasgos (AMR) y Análisis Multidimensiones (AMD) fue creado originalmente como un método analítico para el estudio detallado de las variaciones entre registros. Ello permite describir y comparar múltiples textos de diversos corpus en estudio. Este enfoque metodológico fue inicialmente desarrollado, según Biber (1988, 1994) para:

– Determinar los patrones lingüísticos sobresalientes y en co-ocurrencia en una lengua, desde una perspectiva empírica cuantitativa, y

– Comparar registros orales y escritos en un espacio lingüístico definido por aquellos patrones en co-ocurrencia.

Un aspecto clave para la descripción profunda de los textos de un corpus es el enfoque comparativo, es decir, construyendo una base de datos de corpus variados y a partir de registros diversificados. Diversos investigadores destacan que la investigación más reciente revela la importancia de emplear la comparación de corpus con registros de naturaleza diversificada para develar diferencias significativas sin importar necesariamente el tamaño de las muestras. Biber (1988) ha sido uno de los pioneros en mostrar la relevancia de la comparación de registros, ya que la descripción de un determinado corpus se enriquece y se hacen evidentes sus rasgos prototípicos por medio de la comparación y el contraste. La mera descripción de un objeto aislado es menos informativa y no logra revelar ciertas características intrínsecas del mismo. Muchas de ellas emergen y adquieren relevancia al ser puestas en contexto, por ejemplo, con otros registros.

Desde este marco, buscando la comparación sistemática, Biber (1988) y Biber, Conrad y Reppen (1998) han dado cuenta de interesantes variaciones sistemáticas de orden gramatical y léxico en diversos registros del inglés oral y escrito. Por su parte, Parodi (2007) también ha dado cuenta, entre otras, de variaciones entre oralidad y escritura para el español. Dos hallazgos entre muchos de los reportados parecen relevantes: por un lado, los rasgos lingüísticos individuales presentan una ocurrencia diversa en variados registros; por otro, los mismos o similares rasgos

lingüísticos pueden tener funciones diferentes al entrar en combinaciones con otros rasgos y, al aparecer, en textos pertenecientes a registros diversos. En este sentido, una de las fortalezas de este enfoque metodológico se funda en un principio lingüístico comunicativo que resulta extremadamente sensato: la variación entre registros no se explica únicamente por un solo parámetro o dimensión, lo que equivale a sostener que existen múltiples distinciones situacionales entre registros. Dicho de otro modo, no es posible pensar que un rasgo lingüístico o, incluso, unos pocos de ellos puedan explicar exclusivamente una determinada variación entre registros (por ejemplo: oral/escrito, formal/informal). Las investigaciones en que se ha aplicado el análisis multivariado han revelado que diferentes dimensiones se construyen a partir de conjuntos diferentes de rasgos lingüísticos co-ocurrentes, reflejando así diversas interpretaciones funcionales subyacentes (por ejemplo: objetividad, abstracción de información, modalización). Del mismo modo, las tradicionales distinciones de índole más dicotómica (interactivo/no-interactivo), se ven desafiadas por los estudios con análisis multidimensional, ya que se ha llegado a demostrar que existe un *continuum* de variación lingüística a lo largo de los registros. Por supuesto, esto último es concordante con las investigaciones que adscriben a la idea de categorías de limites difusos (*fuzzy categories*) y que hoy en día tienen gran aceptación entre la comunidad científica.

3.2. Point de vue de Jean-Marie Pierrel

Les besoins de codage standardisé de corpus abordés déjà au point 1 sont importants dans ce cadre. Que cela soit pour des corpus comparables (en termes de genres, époques, langues, etc.) ou pour des corpus parallèles multilingues, dont on imagine volontiers qu'ils ne peuvent être élaborés par un chercheur seul ou même par une seule équipe, les conditions de mutualisation des ressources prennent ici une importance toute particulière. Trois points méritent à ce niveau toute notre attention :

– les formats

– les référentiels de vocabulaires utilisés

– la mise en place d'une interconnexion simple entre nos entrepôts de données.

3.2.1. Concernant les formats

Nous rappellerons ici simplement la nécessité qu'il y a d'utiliser des formats standard (XML pour le langage de base, TEI pour la définition des schémas d'annotation, UTF8 pour le codage des caractères par exemple). A cela s'ajoute bien sûr la nécessité de disposer de métadonnées précises permettant de caractériser les corpus mis à disposition de la communauté. Au-delà de simples métadonnées de type bibliographiques, il importe de préciser dans ces métadonnées :

– la langue (ou les langues dans le cas de corpus parallèles)

– la date ou l'époque d'écriture ou de production de la ressource primaire et pas uniquement l'époque d'édition du document. À titre d'exemple l'édition critique de la première traduction en français par Raoul de Presles de *La Cité de Dieu* de saint Augustin (5 volumes contenant les 22 livres de l'œuvre) réalisée par l'équipe du projet ERC animé par notre collègue Olivier Bertrand doit clairement indiquer la date originale de cette traduction 1371-1375 et non pas uniquement sa date de publication qui sera dans ce cas 2013.

– le genre, etc.

Mais cela ne suffit pas. Encore faut-il pouvoir unifier ou déclarer des vocabulaires de données issus si possible de référentiels communs.

3.2.2. Concernant les référentiels de vocabulaires utilisés

Lorsqu'on définit des métadonnées (genres, langues, etc.) ou des annotations sur des données primaires (annotations morphosyntaxique, parties de discours...) et afin de faciliter leur réutilisation par d'autres, il apparait indispensable d'utiliser un vocabulaire contrôlé et de le référencer, voire mieux, d'utiliser un vocabulaire contrôlé existant. Là encore diverses initiatives internationales vont en ce sens en proposant des registres de catégories de données. ISOcat[7] en est un bon exemple. ISOcat est un registre central de catégories de données pour tous les concepts pertinents en linguistique et dans le domaine des ressources linguistiques, y compris les catégories de métadonnées. L'interface Web ISOcat[8] permet de rechercher des catégories de données qui vous intéressent, de sélectionner celles

[7] www.isocat.org

[8] https://catalog.clarin.eu/isocat/interface/index.html

que vous souhaitez utiliser, et, dans le cas où aucune catégorie de données ne correspond à vos besoins, de définir et proposer votre catégorie de données dans ISOcat.

Là encore il convient de s'organiser pour que la communauté des linguistes et philologues de la Romanie soit plus présente dans les groupes de travail à l'initiative de ces standardisations. La meilleure standardisation dans nos domaines sera en effet celle que nous définirons ensemble !

3.3. Point de vue d'Achim Stein

Malgré les efforts de standardisation sur le plan formel (XML, TEI etc.), la comparaison de corpus se heurtera toujours, et nécessairement, aux différents types d'informations linguistiques présentes dans les corpus. Ces différences sont, à mon avis, non seulement inévitables, mais aussi nécessaires si les corpus sont censés refléter l'état de l'art et le progrès linguistique. Elles sont causées d'une part par les catégories spécifiques qui sont nécessaires pour annoter certaines variétés (diachroniques, diastratiques, diamésiques, etc.) et de l'autre par les différents modèles linguistiques qui sont mis en œuvre dans l'analyse des corpus.

La deuxième cause me semble incontournable. La plupart des corpus sont créés pour faire avancer ou pour faciliter la recherche linguistique dans un certain domaine. L'objectif primaire n'étant alors pas la couche « texte », mais la couche « annotation », il est normal que cette dernière reflète les objectifs et par conséquent les modèles linguistiques du projet. La déclaration des catégories sur ISOCAT (mentionnée par J.-M. Pierrel), est une mesure importante et nécessaire, mais qui devra sans doute être étendue, puisque les couches d'annotation plus « élevées » (syntaxe, sémantique) reposent sur des interactions complexes et ne peuvent être que difficilement ramenées à la définition de catégories isolées.

Tant que les théories linguistiques évolueront (ce que j'espère vivement), nous ne pourrons de toute façon pas éviter le problème de l'hétérogénéité des corpus. Mais nous pourrions l'amoindrir, par exemple en publiant systématiquement non seulement une documentation cohérente du modèle, mais aussi des corpus vérifiés (*gold standard*), facilitant la réutilisation d'un modèle ou la création d'outils de conversion.

3.4. Point de vue d'Andrea Bozzi

Il problema di una rappresentazione standardizzata dei fenomeni linguistici è di vitale importanza anche per la creazione di corpora paralleli multilingui. Per esempio, la codifica dei valori grammaticali (noti col nome di POS, 'part of speech') attribuiti da un sistema manuale, automatico o semiautomatico (ad esempio, un analizzatore morfologico, un sistema di lemmatizzazione, un 'parser' sintattico, un 'disambiguatore' semantico per i casi di omografia fra forme identiche, ma appartenenti a lemmi differenti, ecc.) può essere ormai eseguita con sistemi di marcatura condivisi. Mi sembra che i tempi siano ormai maturi, visti i risultati sempre più positivi ottenuti nell'ambito di progetti internazionali nel campo della Linguistica Computazionale, per diffondere maggiormente le informazioni sulle tecniche, i metodi e le procedure che hanno consentito progressi così consistenti affinché siano utilizzati anche dai cultori delle discipline filologiche i quali, in linea di massima, ne hanno informazioni solo di carattere generale. Mi permetto di ricordare che alcuni programmi realizzati, appunto, nell'ambito della Linguistica Computazionale sono stati originati dalla necessità di costituire corpora multilingui paralleli: i primi risalgono addirittura al periodo in cui la Commissione Europea finanziò ricerche nell'ambito del progetto EUROTRA (fine anni '70) per la realizzazione di programmi di traduzione automatica fra lingue comunitarie (in un primo tempo 7 e successivamente 9).

Questa iniziativa ultradecennale, anche se non fu in grado di produrre molti dei risultati attesi, ebbe tuttavia il merito di dare avvio a corpora paralleli la cui annotazione multilivello (soprattutto morfologica, sintattica e lessicale) era propedeutica alla preparazione dei programmi di traduzione vera e propria. L'evoluzione tecnologica, sia per quanto riguarda il software che l'hardware, ha reso i sistemi di allineamento fra testi paralleli in varie lingue molto più efficienti e, soprattutto, veloci: i metodi statistici e stocastici possono venire applicati su quantità di dati incomparabilmente superiori a quelli sui quali si operava negli anni '80, con la conseguenza di poter disporre di risultati quantitativamente e qualitativamente determinanti per lo studio della variazione fra una lingua ed un'altra. Ne hanno tratto vantaggio, soprattutto, le valutazioni di carattere morfosintattico e semantico. Ritengo che le strategie e le procedure oggi disponibili possano venire sperimentalmente applicate anche a corpora di lingue romanze di epoca medievale per avere a disposizione dati di comparazione concreti e in numero sufficiente a confermare, o eventualmente rivedere, la spiegazione di fenomeni evolutivi rispetto al latino, soprattutto quando si tratti di divergenze attribuibili a variazioni geolinguistiche o a condizionamenti socioculturali. Tali corpora e le annotazioni eseguite con l'aiuto di sistemi computerizzati consentono, inoltre, di indagare meglio anche le varietà linguistiche nell'ambito di un medesimo idioma per comprendere il rapporto sincronico e diacronico che lega un sistema

linguistico alle proprie variazioni (come, per esempio, le variazioni dialettali o quelle intervenute, come si è detto, per cause socio-culturali).

4. L'édition de textes et de documents sonores

L'édition de textes (des codex médiévaux au corpus universel de la toile d'araignée mondiale WWW) et de documents sonores (de l'enquête sur le terrain traditionnelle au foisonnement de la production audiovisuelle) : que retenir et comment traiter ? L'établissement de lexiques de ces documents.

4.1. Point de vue de Jean-Marie Pierrel

L'apport de la linguistique de corpus à la compréhension des phénomènes langagiers est aujourd'hui devenu fondamental. Le nombre d'énoncés qu'entend et produit une personne durant sa vie est très grand. Grâce à l'augmentation de la variété et de la taille des corpus, il est aujourd'hui devenu possible de démontrer les faits langagiers à l'aide d'exemples attestés en grand nombre et de tester les propositions de la linguistique. Pour cela, un grand nombre de corpus contrôlés, bien décrits et variés, est nécessaire.

Non spécialiste de l'édition ni de textes écrits ni de documents sonores, je voudrais ici à Nancy rappeler les nombreuses contributions importantes à la lexicographie française qui ont vu le jour grâce à une exploitation raisonnée de corpus informatisés à Nancy, au sein du Centre de Recherche pour un Trésor de la Langue Française (CRTLF), puis de l'Institut National de la Langue Française (INaLF) et enfin de notre laboratoire qui[9] se veut être aujourd'hui le digne successeur de ces deux structures.

Initiées au départ autour du projet de *Trésor de la Langue française* (TLF), premier dictionnaire de langue se fondant sur une méthodologie systématique d'analyse des usages effectifs des mots de notre langue à travers l'exploitation d'une vaste base de données textuelles dont la saisie a débuté dès les années 60, ces études nancéiennes en lexicographie française se sont poursuivies, au-delà de la rédaction du TLF, suivant deux orientations complémentaires : la lexicographie historique, fortement appuyée sur la notion de corpus, et la valorisation informatique des ressources produites. Ces orientations ont provoqué sur le plan des études lexicales une véritable révolution qui fit de l'informatique un outil indispensable pour :

– étudier le lexique et ses propriétés à travers l'exploitation intelligente de textes et de documents ;

– structurer et normaliser les connaissances lexicales et lexicographiques ;

– valoriser, partager et mutualiser les résultats de la recherche sur le lexique de notre langue, trop souvent encore dispersés.

Rappelons à titre d'exemple quelques-unes de ces contributions : le *Trésor de la Langue Française Informatisé* ou TLFi (Dendien et Pierrel 2003) qui correspond à la valorisation informatique, sous forme de CDRom et sur le Web (www.atilf.fr/tlfi), du dictionnaire de référence qu'est le TLF. Au-delà d'une simple informatisation de la version papier du TLF, ce projet a permis de poser les bases d'une exploration nouvelle de données lexicographiques autorisant des parcours vraiment novateurs de ces matériaux et de donner une seconde vie à ce grand dictionnaire qui aujourd'hui, à n'en point douter, est le dictionnaire institutionnel français le plus consulté sur le Web. Le *Dictionnaire du Moyen Français* ou DMF (Martin, Gerner et Souvay 2007 ; Städtler 2010 : www.atilf.fr/dmf) qui articule fortement valorisation informatique et lexicographie historique. Avec comme objectif de combler une lacune existante entre le Tobler-Lommatzsch en amont et le Huguet en aval, ce dictionnaire se propose de décrire le lexique français de la période 1330 – 1500 et s'appuie sur un concept de lexicographie évolutive (Martin 2008) que l'informatique autorise aujourd'hui. D'autres développements récents en lexicographie historique avec en particulier le programme TLF-Etym (Steinfeld et Andronache 2011 : www.atilf.fr/tlf-etym) qui se propose de réviser progressivement les notices intitulées « Etymologie et histoire » du TLF ; la Base des Mots-fantômes (Steinfeld 2013 : www.atilf.fr/MotsFantomes), ressource de métalexicographie critique dont l'ambition est de recenser ces pseudo-lexèmes disposant à tort d'un statut lexicographique (« ces mots qui n'existent pas »), les sens fantômes et les lemmatisations erronées qui se trouvent dans les dictionnaires historiques et étymologiques français de référence, mais dont on ne trouve aucune attestation véritable dans les grands fonds de textes historiques numérisés. Enfin le Portail lexical du Centre National de Ressources Textuelles et Lexicales (CNRTL : www.cnrtl.fr/portail) créé par le CNRS au

[9] Analyse et Traitement Informatique de la Langue Française : www.atilf.fr

sein de notre laboratoire. Ce portail lexical a pour vocation de valoriser et de partager, en priorité avec la communauté scientifique, un ensemble de données issues des travaux de recherche sur le lexique français. Projet évolutif, cette base de connaissances lexicales propose, à partir d'une unité lexicale, d'intégrer un maximum de connaissances disponibles.

Comme nous l'avons montré dans (Pierrel et Buchi 2009), la mise à disposition très large des ressources linguistiques du DMF a permis des avancées scientifiques dans des domaines non entrevus lors de la définition de cette ressource, par exemple la pragmatique et la morphologie constructionnelle. C'est pourquoi à la question souvent posée de savoir quelles ressources retenir pour une mise à disposition de l'ensemble de la communauté scientifique, ma réponse est simple : toute ressource définie et construite dans le respect d'une bonne démarche linguistique et philologique justifie une telle diffusion.

4.2. Point de vue d'Achim Stein

En abordant les questions d'édition, nous nous plaçons dans un domaine plus large, non exclusivement réservé aux linguistes. Car souvent, les corpus proviennent d'autres domaines (littérature, histoire, sociologie, médias, etc.) ou bien ils sont d'origine linguistique, mais intéressent ces autres domaines. Ici, nous retrouvons donc le sujet de l'annotation multicouche, et des principes favorisant une annotation techniquement indépendante de la ressource (*stand-off annotation*).

Mais la conséquence la plus importante de cette ouverture vers d'autres domaines est, à mon avis, une contrainte qui s'impose à nous, linguistes ou créateurs de corpus. Cette contrainte dit que tout utilisateur d'un corpus, linguiste ou non, doit être capable d'évaluer le texte et son annotation de manière réfléchie et critique. Et les conséquences concernent bien entendu non seulement l'édition du texte, mais aussi son traitement ultérieur, c'est-à-dire les différents types d'annotation qui s'y ajoutent ainsi que les ressources externes (lexicales ou autres) qui se greffent sur ce texte.

Deux exemples :

– Représenter la fiabilité de l'annotation, dans l'annotation même ou par les outils qui servent à exploiter ce type d'annotation. Concrètement, un utilisateur qui obtient le résultat d'une requête qui repose sur certaines catégories, morphologiques ou syntaxiques, devrait être conscient du taux d'erreur de cette requête. Dans le meilleur des cas l'outil d'exploitation est informé sur les taux d'erreur des outils d'annotation, parce que ces derniers, justement, les ont ajoutés à la ressource, parallèlement à la déclaration des catégories, dans l'en-tête du document, par exemple.

– Représenter les ambiguïtés d'analyse dans l'annotation. Le problème ressemble au précèdent. Les outils d'annotation peuvent réduire le nombre des décisions prises lors de l'analyse en annotant les alternatives. Ceci augmente le « bruit » dans le résultat d'une requête et pose des problèmes de représentation aux standards d'annotation.

Bref, tout comme l'éditeur d'une édition critique documente les décisions qu'il prend et les problèmes qu'il rencontre, l'annotateur d'un corpus, humain ou machine, doit, lui aussi, documenter ses décisions et les intégrer, si possible, dans la ressource.

4.3. Point de vue d'Andrea Bozzi

Lo sviluppo della tecnologia digitale e del Web impone di porre l'attenzione su due aspetti:

– esiste una differenza fra: a) edizione elettronica mediante strumenti ipertestuali/multimediali e: b) edizione assistita da calcolatore mediante applicazioni Web anche di tipo collaborativo;

– esistono ormai molti modi diversi di produrre il lessico dei documenti digitali (dall''index lemmatum' all'indice ontologico-concettuale sulla base delle tecniche del 'semantic web') e delle modalità con cui si strutturano le descrizioni delle singole voci (dalla descrizione non strutturata alla descrizione secondo i metodi del lessico generativo computazionale) [10].

Per quanto riguarda il primo punto, la marcatura da parte dell'editore critico di informazioni di tipo diverso (il testo, l'immagine digitale della fonte, i dati di apparato, la bibliografia, gli eventuali passi paralleli, le note critiche, ecc.) mediante l'utilizzo di 'link' ipertestuali produce edizioni elettroniche (sul Web o su supporti ottici) ove il ruolo

[10] Si veda Pustejovsky, James (1995). Una interessante applicazione di questo metodo al lessico di Ferdinand de Saussure si legge in Ruimy, Nilda / Piccini, Silvia / Giovannetti, Emiliano (2012, 1043-1056).

del sistema informatico è piuttosto banale e limitato. Esso, infatti, non fa altro che valutare le marcature e offrire all'utilizzatore una lettura dei dati secondo la sequenza con la quale egli intende 'navigarli' (dal testo all'immagine, oppure dal testo all'apparato delle varianti, oppure ancora dalle varianti alla bibliografia, ecc.). L'edizione assistita da calcolatore, invece, è definibile come filologia computazionale in senso stretto poiché, su richiesta dello studioso, la macchina interviene attivamente nel processo critico-editoriale mettendo in moto componenti software tanto più numerosi quanto più completa è l'applicazione che li comprende. Essi, cooperando come moduli interrelati, producono risultati mirati e quantitativamente rilevanti grazie ai quali si mette in condizioni lo studioso di effettuare interventi critici più sicuri e documentati. Tale approccio facilita la registrazione di apparati, la costituzione del testo e, grazie a tecniche di 'clustering' e di 'information visualisation', anche la produzione di ipotesi stemmatiche.[11]

Per quanto riguarda il lessico, anche in questo caso i sistemi di elaborazione elettronica dei dati testuali hanno da decenni sviluppato programmi molto efficienti di indicizzazione per la produzione di liste alfabetiche delle forme. Essi, a loro volta, sono stati completati da moduli di analisi morfologica in grado di attribuire automaticamente le forme flesse ai lemmi, spesso con informazioni sufficienti per la disambiguazione delle forme omografe derivate da lemmi etimologicamente diversi. Si sono in tal modo resi disponibili sia 'indices verborum' di autori o di interi 'corpora', ma anche 'indices lemmatum' utilissimi alla preparazione di studi e ricerche lessicali. Un aspetto correlato con quest'ultimo consiste nella descrizione dei valori semantici che un lemma possiede nei testi del corpus o, in ultima analisi, nel dizionario generale della lingua di riferimento. Nel corso delle ricerche in Linguistica Computazionale si sono susseguiti vari criteri adatti a tale scopo, ma quello che sembra offrire ottime capacità descrittive, indipendentemente dalla lingua al quale esso si applica, si basa sulla teoria del lessico generativo. Essa appare molto più versatile del noto WordNet (ItalWordNet per l'Italiano e EuroWordNet per le relazioni fra elementi semantici fra più lingue europee, mediante il parallelo fra gruppi di sinonimi)[12]. La connessione fra il lessico semanticamente organizzato e l'archivio dei testi, poi, consente di effettuare ricerche molto dettagliate e mirate al ritrovamento solo di quei passi che documentano un valore semantico specifico, escludendone altri.

4.4. Point de vue de Giovanni Parodi

¿Qué mantener en el registro de un análisis de corpus? y ¿cómo procesar automática o manualmente textos de naturaleza diversa? Ambas preguntas conllevan desafíos y proyecciones para el tratamiento de corpus e imponen una agenda programática a la que ciertamente se debe atender.

A la primera pregunta ya se abordaba en los puntos precedentes, pues un asunto relevante es de qué tipo de marcaje se puede disponer para un estudio de corpus. Clásicamente los etiquetadores morfosintácticos han estado en el foco y su adecuado desempeño sigue siendo materia de investigación y desarrollos. Alcanzar un 100% de marcaje fiable es un desafío. Otros tipos de marcas más textuales y más discursivas también estan siendo estudiadas y constituyen núcleo de atención. Por otro lado, también considerábamos que información extratextual de tipo contextual es relevante y debe ser codificada: tal como procedencia, fecha de recolección, lugar, etc. También información de orden temático, tipo de género discursivo, registro, etc.

En cuanto al procesamiento computacional de textos de variada naturaleza se ha abierto un campo de proporciones tremendo. Por una parte, el análisis de textos orales impone retos muy motivadores: dar cuenta de los rasgos de la oralidad y su contexto de producción (rasgos de índole pragmática) es un terreno en franco desarrollo. Por otra parte, el denominado re-descubrimiento de lo obvio, es decir, la característica multimodal de todo texto y no solo su constitución a partir de palabras (sistema verbal), sino de múltiples sistemas semióticos (gráfico, matemático, tipográfico, entre otros), ha creado un ámbito de investigación muy promisorio (Parodi, 2010b). Incluso ya no solo se abordan los denominados *textos estáticos* (en formato papel o tipo PDF en computador), sino que los *textos dinámicos* (cine, video, hipertextos, textos web, etc.) ofrecen un área de estudio innovadora.

[11] Un esperimento in tal senso è descritto in Corradini (2005, 355-368). Più di recente ha sviluppato tecniche di stemmatica computazionale Camps, Jean-Baptiste (2013a e 2013b).

[12] Si vedano il sito ufficiale di WordNet (<www.illc.uva.nl/EuroWordNet>) e Rodrìguez, Horacio / Climent, Salvador / Vossen Piek. et alii (1998, 117-152).

5. La pérennisation des documents

La pérennisation des documents : comment conserver, mais aussi diffuser le plus librement possible un maximum de documents soigneusement traités ; rôle des diverses institutions pour l'archivage des corpus, leur transcodage régulier, leur mise à la disposition du public au fil des ans.

5.1. Point de vue d'Achim Stein

Je ne répéterai pas l'importance des grandes initiatives européennes, DARIAH, CLARIN etc., qu'A. Bozzi et J.-M. Pierrel ont évoquée. Mais il serait important de signaler plus clairement les liens entre la Société de Linguistique Romane et ces initiatives, en tant qu'institution ou à travers les projets de recherche menés par ses membres. Une fois de plus, je vais distinguer entre la pérennisation des documents statiques d'une part, qui concerne la philologie et dont mes collègues reparleront, et les outils et les méthodes, dynamiques et davantage liés à la recherche linguistique, de l'autre. Concernant ce deuxième point, je rappelle que le projet européen CLARIN crée des plateformes qui permettent de traiter – baliser et annoter – les corpus en ligne, sur des serveurs décentralisés. Les outils de traitement sont donc disponibles sur les serveurs qui permettent aux utilisateurs de créer des chaînes de traitement.

L'avantage de ce concept est tout d'abord que les outils sont ainsi accessibles à tout le monde, et qu'ils ne nécessitent aucune installation technique. Le deuxième effet, plus important à long terme, est le fait que la longévité des outils et des méthodes est ainsi garantie. A mon avis, c'est cette facilité d'accès qui va produire un effet de standardisation, dans la mesure où elle permettra au « grand public » de produire des textes pourvus d'une annotation de qualité.

5.2. Point de vue d'Andrea Bozzi

Nel corso degli ultimi anni si è in più occasioni lamentata l'incertezza della permanenza dei dati sui supporti fisici adottati e si sono levate critiche sulla relativa tecnologia (prima magnetica, come ricorderanno tutti coloro che hanno lavorato con i nastri sui quali erano registrate le informazioni, poi digitale). Secondo la mia personale esperienza queste osservazioni non sono sempre giustificate poiché lavori che io stesso ho eseguito alla fine degli anni '70, opportunamente riversati sui nuovi mezzi di archiviazione mano a mano che essi erano immessi sul mercato e venivano acquisiti nei grandi centri di elaborazione dati, sono ancora integri e su essi possono essere applicati i nuovi e più potenti sistemi di trattamento e di analisi.

Un altro aspetto, piuttosto, deve essere preso in considerazione e riguarda la riconversione degli elementi di marcatura affinché da quelli originari, spesso sviluppati e utilizzati da un singolo, si possa ottenere una ri-codificazione dei testi secondo metodi standardizzati o, almeno, compatibili con gli standard condivisi a livello internazionale. Questo aspetto ci riconduce a quanto ho detto relativamente al primo tema trattato e al quale io attribuisco un grande valore. La condivisione di sistemi di codifica, soprattutto di fronte all'affermazione indiscussa del Web come infrastruttura comunicativa mondiale, offre l'opportunità alle Scienze Umane e alle comunità di ricercatori che in esse operano di mettere a frutto conoscenze, dati, programmi, risultati troppo spesso isolati gli uni rispetto agli altri. Le Scienze umanistiche digitali potranno avere molti più elementi di aggregazione rispetto al passato e assumeranno un modo di agire simile a quello che caratterizza altri domini scientifici (in particolare, la fisica e la genetica). Affinché questa mia osservazione non rimanga vaga, ripeto in questa occasione un esempio che ho altre volte citato. La scelta di una lezione per stabilire un testo occitano medievale di argomento medico farmaceutico potrebbe avvalersi del raffronto con vasti archivi di lingue romanze (testi e dizionari) consultabili sul Web per verificare attestazioni o interrogare eventuali studi e saggi monografici. Se queste operazioni di raffronto sono sempre state fatte da studiosi scrupolosi ed attenti sugli strumenti cartacei disponibili, oggi l'indagine può (o potrebbe) essere condotta in maniera molto più agevole e rapida solo se i dati sono resi accessibili in linea e se essi hanno una esplicita e condivisa rappresentazione nella rete. Per questa ragione sento il dovere di complimentarmi vivamente con il lavoro svolto a Nancy dall'ATILF che ospita questo convegno per il grande lavoro svolto proprio in questa direzione e plaudo a tutte le iniziative di valore scientifico che seguono questo esempio virtuoso, grazie agli sforzi di ricercatori di tutto il mondo.

Per questa serie di motivi l''open data' e, per quanto riguarda i sistemi e le applicazioni computazionali, la loro disponibilità 'open source' senza pagamento di diritti di uso o riproduzione, è la reale prospettiva del futuro, soprattutto per i lavori che sono stati condotti da studiosi sulla base di finanziamenti pubblici. Le Istituzioni pubbliche e le società scientifiche hanno, pertanto, il compito di:

– garantire i 'credits' e l''authorship' della documentazione (edizioni, sistemi, lessici, ecc.) immessa in rete su portali istituzionali e a questa documentazione deve essere attribuita adeguata validità come prodotto della ricerca valutabile;

– garantire la transcodifica dei valori propri di un qualunque sistema proprietario verso un sistema standard fra i quali quello che si è affermato, nel settore degli studi del testo, è la TEI;

– garantire una maggiore attenzione rispetto a quanto non è avvenuto fino ad oggi verso le infrastrutture di ricerca invitando a tenere in considerazione, nelle fasi di progettazione, anche le esigenze dei filologi oltre a quelle, oggi prioritarie, dei linguisti che operano su lingue vive moderne, degli storici, dei sociologi, ecc.

5.3. Point de vue de Giovanni Parodi

Para la adecuada permanencia y disponibilidad de los corpus se hace fundamental contar con interfaces computacionales en línea, construidas por medio de sitios web. Si los corpus son accesibles libremente en línea y se les puede consultar en formato "marcado", se asegura su perpetuidad y se garantiza que el trabajo realizado sea productivo. Lamentablemente aún existen corpus que tienen carácter privado o cuyo acceso no es gratuito. O, incluso, que se recolectan pero nunca logran estar en un sistema abierto que permita el libre acceso y consulta.

Es evidente que existen costos económicos asociados a esta disponibilidad en línea. Es de esperar que instituciones universitarias y de investigación logren apoyar estos proyectos y se vaya paulatinamente haciendo más factible el acceso gratuito en línea.

En el caso de la lengua española, diversas universidades e instituciones gubernamentales de apoyo a la investigación han respaldado la creación de sitios web que alojan y permiten el acceso a corpus digitales. También se han creado interfaces de consulta que permiten a los investigadores obtener datos estadísticos de los textos analizados así como descripciones de estos corpus. La labor de la Real Academia Española (RAE) ha sido significativa en este sentido en los últimos años. La RAE mantiene acceso en línea y de modo gratuito de importantes corpus (www.rae.es). Lo mismo se puede decir del IULA de la Universidad Pompeu Fabra y otras tantas universidades españolas.

En el caso de Chile, una importante tarea en la investigación de corpus ha venido realizando la *Escuela Lingüística de Valparaíso* (ELV: www.elv.cl) de la Pontificia Universidad Católica de Valparaíso con la creación de El Grial (www.elgrial.cl). El Grial es tanto una herramienta de etiqueta morfosintáctico para textos en español como también una plataforma en web de consulta de corpus textuales. En esta plataforma se obtiene información detallada de cada uno de los corpus disponibles gratuitamente en línea. Una de sus muchas fortalezas es el acceso a corpus diversificados, es decir, de disciplinas diversas, y separados, entre otros, por géneros del discurso y temáticas.

5.4. Point de vue de Jean-Marie Pierrel

L'importance d'un plus grand partage de nos ressources linguistiques, en particulier de nos corpus, au sein de la communauté de linguistique et philologie romanes, plaide pour la mise en place d'entrepôts de données fiables. En France, nous venons de bénéficier d'un soutien conséquent (2,6 millions d'euros HT) dans le cadre du Programme d'Investissement d'Avenir (PIA) de l'État français pour construire un équipement d'excellence ORTOLANG (Open Resources and Tools for LANGuage, www.ortolang.fr) dont nous assurons la responsabilité à Nancy.

Une telle infrastructure a pour but de proposer un réservoir fiable de données (corpus, lexiques, dictionnaires, etc.) et d'outils sur la langue et son traitement clairement disponibles et documentés qui valorise le français et les langues de France à travers un partage des connaissances sur notre langue accumulées par les laboratoires publics. Un tel équipement a aussi pour objectif de généraliser et d'assurer la pérennisation des efforts entrepris par les Centres de Ressources Numériques sur la langue mis en place par le CNRS (Pierrel et Petitjean 2007) : CNRTL (Centre de Ressources Textuelles et Lexicales www.cnrtl.fr) et SLDR (Speech and Language Data Repository, http://sldr.org, anciennement CRDO-Aix). Enfin, il servira de plateforme technique sur la langue, écrite et orale, support des actions de coordination menées par la grande Infrastructure de recherche française Huma-Num (http://www.huma-num.fr/), fusion d'ADONIS et de CORPUS, dans laquelle nos laboratoires et centres ressources sont fortement impliqués. ORTOLANG se propose donc de mettre en place un processus permettant à une donnée linguistique, une fois créée, d'être cataloguée, éventuellement améliorée (voire corrigée), puis diffusée et enfin archivée. Le modèle d'ORTOLANG reprend les entités de base du modèle OAIS en précisant le cycle de correction / enrichissement des données, rendu possible par l'archivage intermédiaire. Les fonctions envisagées pour cet équipement sont donc :

– Permettre, au travers d'une véritable mutualisation, à la recherche sur l'analyse, la modélisation et le traitement automatique de notre langue de se hisser au meilleur niveau international.

– Aider à la création de données : faciliter l'accès à des instruments permettant l'acquisition ou la création de données (par exemple, numérisation, chambre sourde, caméra rapide, mouvements oculaires, électro-encéphalographie, articulographe, etc.).

– Enrichir les données : plusieurs outils permettent d'enrichir automatiquement les données brutes (étiquetage morphosyntaxique, analyses prosodiques, syntaxiques, etc.) seront proposés sur la plateforme ORTOLANG.

– Accompagner les auteurs sur les standards, les normes et les recommandations internationales actuelles : XML, TEI, LMF, MAF et SYNAF ; enrichissement de ressources et des outils.

– Identifier les données : catalogage des ressources et outils existants à travers un ensemble de métadonnées normalisées ; contrôle et validation des ressources et des outils.

– Assurer la pérennité des données au travers d'une double fonction de stockage sécurisé etd'archivage pérenne : stockage, maintenance et curation des ressources et des outils ; archivage pérenne, à travers la solution mise en place par la TGIR Huma-Num en lien avec le CINES.

– Assurer la diffusion : aide et accompagnement des utilisateurs et mise en place des procédures permettant à des utilisateurs de la plateforme d'exploiter les ressources et outils mutualisés sans avoir à se soucier de leur localisation et implantation géographiques.

Au niveau européen, au cours des dernières années, diverses initiatives ont vu le jour pour permettre une concertation et une mise en réseau d'entrepôts fiables de données. On peut signaler entre autres les deux infrastructures de recherche que sont DARIAH (Digital Research Infrastructure for the Arts and the Humanities, www.dariah.eu) et CLARIN (Common LAnguage Resources and Technology Infrastructure, www.clarin.eu). Il paraît indispensable que nos communautés s'impliquent plus dans ces initiatives.

Il serait en particulier souhaitable de pouvoir disposer d'un véritable réseau d'entrepôt de données de ce type regroupant les diverses langues romanes. L'initiative lancée à l'occasion de ce congrès par notre collègue Antonio Pioletti, président de la Société italienne de philologie romane, de lancer un projet de Conseil international d'Études romanes, qui réunirait les principales associations scientifiques actives dans ce domaine, peut être l'occasion de progresser vers cet objectif. Nous plaidons pour que, dans le programme d'un tel Conseil international d'Études romanes, émerge, entre autres, un projet de mise en place d'un réseau numérique regroupant nos ressources et connecté aux réseaux européens tels que celui que propose CLARIN (Wittenburg 2010).

Pour notre part, en nous appuyant sur l'Equipex ORTOLANG que nous animons depuis Nancy, nous proposons, si la communauté juge intéressante une telle proposition, de participer activement à une telle initiative.

6. Les études linguistiques portant sur les langues romanes

Les études linguistiques portant sur les langues romanes : importance des corpus multilingues pour les études contrastives dans le domaine roman, en particulier pour des branches de notre discipline encore peu développées. Rôle de notre Société pour une meilleure coordination des études romanes.

6.1. Point de vue d'Andrea Bozzi

A costo di correre il rischio di ripetere valutazioni che ho già espresso nel corso del dibattito sugli altri temi della tavola rotonda, credo con convinzione che gli studi sulle lingue romanze possono trarre un grande vantaggio se i loro cultori siano disposti a prestare molta attenzione anche a quanto avviene almeno in altri due settori del 'Computing in the Humanities': la Linguistica Computazionale e la Filologia Computazionale.

La nostra Società presta certamente attenzione all'innovazione tecnologica e nei Convegni Internazionali ha da tempo inserito sessioni che coprono i nuovi strumenti, le nuove tecnologie, le basi di dati a supporto della ricerca linguistica e filologica, i programmi di elaborazione che sono usati con successo in ambito accademico. Tuttavia si dovrebbe forse fare qualche cosa di più per invitare i responsabili dell'educazione universitaria almeno dei Paesi della Comunità a favorire l'inserimento, nel curriculum di studi di linguistica e filologia romanza, di corsi a maggiore carattere tecnologico ove la linguistica e la filologia si possano arricchire dei contributi messi ormai a disposizione dalla scienza dell'informazione, dagli studi statistici, dalle metodologie di annotazione e codifica, dall'elaborazione dei dati. Con questo non intendo affatto sostenere che lo statuto delle nostre discipline debba cambiare o essere totalmente

modificato, ma un aggiornamento si impone per non vederle agonizzare in un mondo in profonda trasformazione e con studenti che vi si dedicano in numero sempre minore.

Quanto ho detto nella discussione circa il terzo tema, per esempio, consente di riproporre metodi di analisi comparativa fra testi e lessici consultabili in linea nell'ambito di un corso universitario che si avvalga di accesso a strumenti computazionali disponibili e approvati scientificamente da parte di un collegio di docenti. Si tratterebbe, quindi, di offrire a docenti e discenti applicazioni grazie alle quali sia possibile fare annotazioni, scrivere piccoli saggi, eseguire analisi linguistiche, trovare citazioni o elementi di intertestualità. La formazione avverrebbe su due binari: quello classico della lezione frontale fra professore e allievi e quello 'strumentale' fra professore, allievi e sistemi informatici pronti ad offrire dati su cui ragionare e registrare i risultati delle osservazioni fatte. L'aspetto partecipativo, che un elemento fondamentale nel Web, avrebbe in tal caso elementi decisamente positivi, anche perché gli allievi potrebbero cooperare alla valutazione dei fenomeni incontrati in un testo anche al di fuori dell'aula, in un ambiente virtuale.

La tecnologia è pronta per tutto questo scenario, ma mancano, probabilmente, alcune condizioni favorevoli la maggior parte delle quali non dipendono direttamente dalle nostre intenzioni: la Società di Linguistica e Filologia Romanza ha però tutto il prestigio internazionale necessario per aiutare, se lo vorrà, a seguire questa direzione, l'unica, a mio avviso, che è in grado di garantire un futuro ai suoi antichi e gloriosi settori di studio. Sono al corrente di quello che avviene in molti nei Paesi europei a proposito degli studi umanistici, ed è di recente pubblicazione su un quotidiano italiano di larghissima diffusione un articolo che lamenta la crisi di iscrizioni al Liceo classico, unica scuola secondaria superiore in Italia che, nell'offerta didattica, abbia lo studio della lingua greca e latina.

Questa tendenza è iniziata alcuni anni fa e non sono pochi i casi nei quali, in alcune città italiane, prestigiosi e storici licei siano stati accorpati ad altre scuole superiori per carenza di iscritti. Come dicevo, le cause sono molteplici, ma un possibile rimedio è l'introduzione di forme di aggiornamento che prevedano l'utilizzo sensato di innovazioni tecnologiche, ancora troppo limitate se non addirittura escluse o adottate sporadicamente e senza un piano formativo che si conformi anche alla presenza di esse in forma organica. Io credo che sia, pertanto, necessario intervenire su due fronti: uno nell'ambito della formazione universitaria ed uno ad un livello più basso. L'innovazione tecnologica e l'adozione di nuovi parametri educativi sono una garanzia di sopravvivenza dal momento che avranno un ruolo determinante nella formazione di nuove generazioni di filologi, dotati di un bagaglio di conoscenze, anche tecniche, che permetteranno alle nostre discipline di venire curate e sviluppate nell'ambito di una cultura che ormai è divenuta digitale.

6.2. Point de vue de Giovanni Parodi

Los estudios de corpus han experimentado un auge tremendo en los últimos años y aunque, como dicen algunos, los corpus siempre han sido parte de la investigación lingüística, nunca como hasta ahora se ha logrado tanta sistematicidad metodológica en torno a objetivos comunes. La expansión de la denominada Lingüística de Corpus en una renovada mirada ha dado pie a magnificas discusiones teóricas y ha abierto escenarios para fortalecer el estudio de la lengua en uso y de la variación en todo sentido; así, los denominados estudios variacionistas que ahora incluyen tanto la variación disciplinar y la de géneros del discurso ha traído atención a aspectos de las lenguas que hasta poco no eran de interés de los investigadores.

La disponibilidad de soporte electrónico para los corpus y el albergarlos en sitios web de acceso abierto y gratuito constituyen oportunidades formidables para el estudio comparativo de las lenguas romances. Ello puede contribuir de modo certero a llevar adelante estudios empíricos basados en corpus que arrojen resultados robustos como nunca antes; también se lograría superar trabajos a partir de textos ejemplares y con escasa proyección, dado lo limitado de algunos corpus. Al mismo tiempo, la existencia de asociaciones como la que acoge esta mesa de discusión con interlocutores de diferentes lenguas y orígenes es una prueba de que se cuenta con interés de colaborar y aportar al desarrollo de la investigación en este campo disciplinar. Así, tanto la LC como los estudios basados en corpus, en un sentido amplio, muestran proyecciones vigorosas.

6.3. Point de vue de Jean-Marie Pierrel

Depuis une dizaine d'années, le paysage de la recherche en linguistique a largement évolué grâce à l'apparition d'importants corpus de langage aisément disponibles sur Internet. Si l'existence d'une linguistique de corpus n'est pas nouvelle, cette évolution de l'accès aux données dynamise de manière très importante le domaine, permet de

démontrer l'importance, du point de vue fondamental, de la notion de variation, et autorise de grandes avancées dans la modélisation des théories exemplaristes ou dites basées sur l'usage.

Si, avant les années 2000, le paradigme générativiste dominait et conduisait à voir les théories et les modèles linguistiques comme fondamentalement sous-déterminés par les données factuelles, ce n'est plus le cas aujourd'hui. Ce sont d'abord les travaux psycholinguistiques d'observation longitudinale, et spécialement ceux menés sur les acquisitions précoces qui ont ébranlé le paradigme cognitiviste chomskyen en documentant une hétérogénéité et une variabilité intrinsèque très importantes et peu compatibles avec l'innéisme de la grammaire universelle. Ces travaux ont récemment rencontré les problématiques de la linguistique variationniste conduites indépendamment depuis plusieurs décennies. La confrontation avec les analyses du changement linguistique en temps réel a par ailleurs souligné l'importance des dynamiques qui structurent, forment et déforment les systèmes linguistiques dans le temps. Enfin, le développement des travaux contrastifs et typologiques a conduit à relativiser la portée des grandes hypothèses universalistes au profit d'une description plus fine et plus précise des données observées. Dans chacun des domaines et des sous-domaines de la linguistique, la notion d'usages ou de pratiques attestés a ainsi été remise au premier plan, induisant un rapport nouveau aux modélisations explicatives et aux formalisations.

Ces théories sont basées sur la notion de constructions, qui sont des associations entre forme et fonction. Les constructions peuvent être extrêmement variées, allant de formes figées (un mot, une holophrase, une expression idiomatique) à des structures plus générales (par exemple la structure transitive sujet-verbe-objet), en passant par de nombreux intermédiaires plus ou moins généralisés (par exemple la construction « c'est X » où « X » peut prendre n'importe quelle forme ; ou la construction « X aime Vinf » où « X » et « Vinf » sont mutuellement contraints). Les constructions peuvent se combiner pour produire des formes langagières de tout niveau de complexité. De telles théories permettent de modéliser la variété à tous les niveaux, de l'interlocuteur à l'intralocuteur. Elles font évoluer le système de catégorisation mis en place sur les exemplaires connus en élargissant sa base empirique, en modifiant le poids fréquentiel d'une série d'exemplaires, en favorisant la formation d'une construction plus générale que celles qui étaient disponibles sous la forme d'exemplaires auparavant.

L'apport de la linguistique de corpus à la compréhension des phénomènes langagiers est donc devenu fondamental. Le nombre d'énoncés qu'entend et produit une personne durant sa vie est très grand. Grâce à l'augmentation de la variété et de la taille des corpus, il est aujourd'hui devenu possible de démontrer les faits langagiers à l'aide d'exemples attestés en grand nombre et de tester les propositions de la linguistique et de la psycholinguistique. Pour cela, un grand nombre de corpus contrôlés, bien décrits et variés, est nécessaire.

De mon point de vue, la Société de Linguistique Romane a un rôle crucial pour accompagner la communauté des chercheurs dans cette évolution fondamentale de nos approches au travers d'initiatives à lancer :

– Pour mieux prendre en compte les traitements informatiques et la linguistique de corpus dans nos formations en linguistique et philologie romanes. En ce sens il convient sans doute de réfléchir pour mieux attirer et accueillir des spécialistes du traitement automatique de la langue au sein des futures éditions de CILPR.

– Pour inciter la communauté de linguistique et philologie romanes à un véritable partage de ressources et d'outils en ce domaine sous le double aspect de l'évolution des mentalités, mais aussi de l'accompagnement à la mise en place de plateformes technologiques de gestion et de mutualisation de telles ressources au sein des diverses communautés linguistiques des langues romanes.

– Pour valoriser cette activité de production des documents et ressources numériques dans l'évaluation des chercheurs et accompagner ou mettre en place des structures servant à la fois de validation et de diffusion de ces productions. C'est en partie du moins le rôle que le CNRS a confié aux Centres Nationaux de Ressources, dont le CNRTL, il y a quelques années et qui en France va être repris par l'équipement d'excellence ORTOLANG que nous animons à Nancy.

6.4. Point de vue d'Achim Stein

Nous vivons une époque de transition où les résultats de la recherche philologique moderne sont publiés sur support numérique. Nous allons vivre une époque où même l'objet de la philologie, le texte primaire, ne sera plus publié ou ne sera plus accessible sur support papier.

Dans les deux domaines qui nous intéressent, la linguistique et la philologie, il faut relever deux défis :

– La linguistique théorique doit se justifier face à un traitement purement automatique des langues qui ne repose plus sur les symboles, ni sur les structures, mais uniquement sur le traitement probabiliste d'un grand nombre de données.

– La philologie historique doit veiller à ce que ses procédures méthodiques s'adaptent, aussi bien que les textes eux-mêmes, aux nouveaux supports.

Je me bornerai à proposer que le CILPR représente de manière plus importante le domaine de la linguistique computationnelle des langues romanes. La section « Travaux en cours » attire telle ou telle contribution issue de ce domaine, mais elle ne suffit pas à rentrer dans le détail des questions méthodologiques dont nous avons souligné l'importance ici.

Certes, mon propos n'est pas de créer un forum pour les linguistes computationnels spécialisés, puisque le CILPR ne saurait de toute façon pas concurrencer les grandes conférences nationales et internationales. Mais ce qui me semble important, ce serait la création d'un niveau d'interface entre ces spécialistes et les philologues en langues romanes.

Concrètement, les travaux computationnels devraient être intégrés d'une manière plus systématique dans les sections thématiques – tout d'abord, parce qu'il est parfois difficile de trancher entre « linguistique » et « computationnel », mais surtout parce qu'il est important d'assurer un maximum de communication entre les deux communautés à l'intérieur des domaines thématiques.

7. Bilan et conclusion (P. Kunstmann)

On a insisté, dans cette table ronde, sur la représentativité des corpus. La standardisation est en grande partie accomplie, même si on peut penser qu'elle ne sera jamais exhaustive. Il convient d'opter pour une mutualisation franche et massive, en respectant toutefois l'éthique scientifique et la simple déontologie.

On a constaté l'utilité de l'analyse à traits ou dimensions multiples, la nécessité d'un vocabulaire contrôlé pour définir les métadonnées, vocabulaire qu'il convient d'enrichir au fur et à mesure de l'évolution des théories et modèles linguistiques. Il importe de diffuser l'information sur les progrès de la linguistique computationnelle auprès des utilisateurs naturels, linguistes et philologues.

Pour les textes statiques ou dynamiques, il faut préciser clairement comment on procède, quelles sont les marges d'erreur des outils de traitement automatique et quel est le pourcentage de bruit auquel on peut raisonnablement s'attendre.

En ce qui concerne l'archivage et la diffusion des données, il semble qu'on avance sur un bon chemin, autant avec le projet ORTOLANG ou le CNRTL à Nancy qu'avec les initiatives de la Real Academia Española ou la plate-forme EL GRIAL pour l'espagnol, par exemple.

Mais tout cela serait à étendre à l'ensemble des langues romanes. Il serait bon que dans les programmes de linguistique de nos universités une large part soit faite au traitement automatique de la langue et que nos futurs congrès, à l'image de celui-ci, s'ouvrent aussi largement à la linguistique computationnelle des langues romanes.

8. Références Bibliographiques

Bernard, Pascale / Lecomte, Josette / Dendien, Jacques / Pierrel Jean-Marie, 2002. « Computerized linguistic resources of the research laboratory ATILF for lexical and textual analysis: Frantext, TLFi, and the software Stella », in : *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas de Gran Canaria, vol 3, 1090-1098.

Biber, Douglas, 1988. *Variation across speech and writing*. Cambridge, Cambridge University Press.

Biber, Douglas, 1994. « Using register-diversified corpora for general language studies », in: Armstrong, Susan (ed.), *Using large corpora*, Cambridge, The MIT Press, 180-201.

Biber, Douglas / Conrad, Susan / Reppen, Randi, 1998. *Corpus linguistics. Investigating language structure and use*. Cambridge, Cambridge University Press.

Biber, Douglas / Reppen, Randi / Clark, Victoria / Walter, Jenia, 2001. « Representing spoken language in university settings: The design and construction of the spoken component of the T2K-SWAL Corpus », in: Simpson, Rita / Swales, John (eds), *Corpus Linguistics in North America*, Ann Arbor, University Michigan Press, 48-57.

Camps, Jean-Baptiste, 2013a. « Detecting Contaminations in Textual Traditions. Computer Assisted and Traditional Methods », presentato all'International Medieval Congress, Leeds, July 3rd 2013 (accessible à l'adresse <www.academia.edu/3825633/Detecting_Contaminations_ in_Textual_Traditions_Computer_Assisted_and_Traditional_Methods>).

Camps, Jean-Baptiste, 2013b. « Sélection des lieux variants et construction d'un stemma : nouvelles expérimentations », presentato a questo stesso XXVII Congrès International de Linguistique et de Philologie Romanes, Nancy 2013.

Corradini, Maria Sofia, 2005. « Formalisation des variantes à des fins computationelles : vérification de l'hypothèse expérimentale sur un texte occitan », in: Buckley, Ann / Billy, Dominique (eds), *Etudes de langue et de littérature médiévales offertes à Peter T. Ricketts,* Turnhout, Brepols, 355-368.

Deerwester, Scott / Dumais, Susan / Furnas, George / Landauer, Thomas / Harshman, Richard, 1990. *Indexing by latent semantic analysis* (accessible à l'adresse <http://www.stanford.edu/class/linguist289/lsi.pdf>).

Dendien, Jacques / Pierrel, Jean-Marie, 2003. « Le Trésor de la Langue Française informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence », *TAL (Traitement Automatique des Langues)*, vol. 44 – n° 2/2003, Hermes Sciences Edition, 11-37.

Dumais, Susan, 1994. *Latent semantic indexing (LSI) and TREC-2* (accessible à l'adresse http://lsi.research.telcordia.com/lsi/LSIpapers.html).

EAGLES, 1996a. *Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora. A common proposal and applications to European languages.* Pisa: ILC-CNR.

EAGLES, 1996b. *Preliminary recommendations on subcategorization* (accessible à l'adresse <http://www.ilc.cnr.it/EAGLES96/synlex/synlex.html>).

Gutiérrez, Rosa María, 2005. « Análisis Semántico Latente: ¿Teoría psicológica del significado? », *Revista Signos. Estudios de Lingüística*, 38 (59), 303-323.

Kunstmann, Pierre, 1982. *Concordance analytique de La Mort le roi Artu*, en collaboration avec M.Dubé, Ottawa, Éditions de l'Université d'Ottawa), 2 vol.

Kunstmann, Pierre, 2012. « L'électronique à l'aide de l'éditeur : miracle ou mirage ? Bilan de quatorze années de travaux au LFA », *Perspectives Médiévales* [En ligne] 34 (*Les textes médiévaux face à l'édition scientifique contemporaine : quels enjeux épistémologiques ?*), 1-51. URL : http://peme.revues.org/2245 ; DOI : 10.4000/peme.2245.

Martin, Robert / Gerner, Hiltrud / Souvay, Gilles, 2007. « Présentation de la seconde version du DMF (*Dictionnaire du Moyen Français*) », in: Iliescu, Maria / Danler, Paul / Siller-Runggaldier, Heidi (eds), *Actes du XXV^e Congrès International de Linguistique Romane (Innsbruck, 3-8 septembre 2007)*, Berlin / New York, Walter de Gruyter, 2010, vol. 6, 213-220.

Martin, Robert, 2008. « Perspectives de la lexicographie informatisée », in: Durand, J. / Habert, B. / Laks B. (eds), *Actes du Congrès Mondial de Linguistique Française - CMLF'08*, ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française, p. 1251-1256, Lexique(s), DOI 10.1051/cmlf08332.

MCVF : *Modéliser le changement : les voies du français*. www.voies.uottawa.ca

Parodi, Giovanni (ed.), 2007. *Working with Spanish corpora*. London / New York, Continuum.

Parodi, Giovanni, 2010a. *Lingüística de corpus: de la teoría a la empiria.* Frankfurt, Veruert/Iberoamericana.

Parodi, Giovanni, 2010b. « Research challenges for corpus cross-linguistics and multimodal texts », *Information Design Journal,* 18 (1), 69-73.

Pierrel, Jean-Marie / Petitjean, Etienne, 2007. « Valorisation et exploitation scientifiques de documents numériques pour la recherche en linguistique : l'exemple du CNRTL », in: *Actes de CIDE 2007 Congrès International sur le Document Numérique*, Nancy, Europia, 13-24.

Pierrel, Jean-Marie / Buchi, Eva, 2009. « Research and Resource Enhancement in French Lexicography: the ATILF Laboratory's computerized resources », in: Bruti, Silvia / Cella, Roberta / Foschi Albert, Marina (eds), *Lexicography in Italy and in Europe*, Newcastle upon Tyne, Cambridge Scholars Publishing, 79-118.

Pivot, Bernard, 2013. *Les tweets sont des chats*, Paris, Albin Michel.

Pustejovsky, James, 1995. *The Generativ Lexicon*, Cambridge, the MIT Press.

Rastier, François, 2004. « Enjeux épistémologiques de la linguistique de corpus », in: Williams, Geoffrey (ed.), *Actes des deuxièmes journées de linguistique de corpus*, (Lorient, 2002), Rennes, Presses Universitaires de Rennes, 31-46.

Rodrìguez, Horacio / Climent, Salvador / Vossen, Piek / Bloksma, Laura / Peters, Wim / Alonge, Antonietta / Bertagna, Francesca / Roventini, Adriana, 1998. « The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology », *Computers and the Humanities*, 32 (2-3), 117-152.

Ruimy, Nilda / Piccini, Silvia / Giovannetti, Emiliano, 2012, « Les Outils Informatiques au Service de la Terminologie Saussurienne », Congrès Mondial de Linguistique Française (Lyon, 2012) (accessible à l'adresse <www.ilc.cnr.it/viewpage.php/sez=ricerca/id=917/vers=ita>), 1043-1056.

Städtler, Thomas, 2010. « Die evolutive Lexikografie am Beispiel der Geschichte des Dictionnaire du Moyen Français », *Zeitschrift für französische Sprache und Literatur*, 120, 1-13.

Steinfeld, Nadine, 2013. « La traque des mots fantômes à travers les terres de La Curne et de Godefroy : un tableau de chasse chargé de trophées pittoresques extraits du *Livre des deduis du roy Modus et de la royne Ratio* », in: Casanova Herrero, Emili / Calvo Rigual, Cesareo (eds). *Actas del XXVI Congreso Internacional de Lingüística y de Filología Románicas* (Valencia, 2010), Berlin / New York, De Gruyter, 7, 411-422.

Steinfeld, Nadine /Andronache, Marta, 2011. « Quoi de neuf du côté de la lexicographie étymologique ? La méthode utilisée dans le cadre du projet TLF-Étym pour distinguer les emprunts au latin de l'Antiquité de ceux faits au latin médiéval », *Estudis romànics* 33, 151-170.

Venegas, René, 2005. *Las relaciones léxico-semánticas en artículos de investigación científica: Una aproximación desde el análisis semántico latente*. Tesis doctoral, Pontificia Universidad Católica de Valparaíso, Chile (accessible à l'adresse http://cybertesis.ucv.cl/tesis/production/pucv/2005/venegas_re/html/index-frames.html).

Venegas, René, 2007. « Clasificación de textos disciplinares en función de su contenido léxico-semántico », *Revista Signos. Estudios de Lingüística,* 40 (63), 239-231.

Wittenburg, Peter / Bel, Nuria / Borin, Lars / Budin, Gerhard / Calzolari, Nicoletta / Hajicova, Eva / Koskenniemi, Kimmo / Lemnitzer, Lothar / Maegaard, Bente / Piasecki, Maciej / Pierrel, Jean-Marie / Piperidis, Stelios / Skadina, Inguna / Tufis, Dan / Veenendaal, Remco van / Váradi, Tamás / Wynne, Martin, 2010, « Resource and Service Centres as the Backbone for a Sustainable Service Infrastructure », in: *Proceeding LREC 2010*, Valetta, Malte, 17-23 May 2010, 60-63.

Pierre KUNSTMANN, professeur émérite à l'Université d'Ottawa

Andrea BOZZI, directeur de l'Istituto di Linguistica Computazionale «Antonio Zampolli», Pise

Giovanni PARODI, professeur à l'Université Pontificale Catholique de Valparaíso

Jean-Marie PIERREL, professeur à l'Université de Lorraine

Achim STEIN, professeur à l'Université de Stuttgart