Claire Clivaz, Jérôme Meizoz,
François Vallotton, Joseph Verheyden (Ed.)
avec la collaboration de Benjamin Bertho

Lire demain

Des manuscrits antiques à l'ère digitale

Reading Tomorrow

From Ancient Manuscripts to the Digital Era

version ebook

Presses polytechniques et universitaires romandes

Interrogation Programme & SuperSenses Extraction : IPSE, une base de données ouverte et flexible

Alberto Roncaccia & Davide Picca – Université de Lausanne

Introduction

IPSE est né à l'Université de Lausanne dans le cadre des séminaires de littérature et d'histoire de la langue italienne avec pour visée d'encourager la pratique de l'interrogation électronique des textes complexes chez des étudiants de niveau Bachelor[1]. D'un point de vue théorique, au cours du projet, nous nous sommes concentrés sur la notion de « texte littéraire » et sur sa perception dans le développement de la Galaxie numérique. Si nous attribuons à la notion de « littéraire » un sens très large, y incluant tout texte à fort coefficient rhétorique et thématique, avec ou sans ambition esthétique, nous pouvons adopter la notion d'« objet littéraire ». L'avantage de cette notion, par rapport à celle de « texte », est celui de pouvoir y inclure l'expérience et la pratique de la collecte d'informations propre au lecteur, en lui reconnaissant le rôle actif de co-énonciateur. Le lecteur produit les critères de cohérence. Ces critères justifient la création d'un corpus de documents qui assume le statut d'« ensemble significatif » au sein d'un « scénario d'interrogation ».

IPSE est basé sur les méthodes de traitement automatique des langues naturelles (TALN). Les démarches de reconnaissance automatique des entités nommées, c'est-à-dire la *Named Entity Recognition* (NER), privilégient trois catégories fondamentales : les personnes physiques, les lieux et l'organisation. Cependant, sur des ensembles de catégories tel que ceux que nous venons d'indiquer, les avis restent très controversés. Comme l'a rappelé Sekine (2004), il

[1] Le projet a été réalisé grâce au Fonds d'innovation pédagogique et au Réseau interfacultaire de soutien enseignement et technologies de l'Unil. Essentielle a été la contribution de Nadia Spang Bovey, ingénieur pédagogique, et de Julien Furrer, developpeur informatique.

faut d'abord établir une catégorisation sémantique du monde pour pouvoir ensuite identifier des catégories pertinentes pour un outil NER. Et en effet, dans la pratique de ce type d'outils, les trois catégories mentionnées (personne, lieu et organisation) sont insuffisantes et génériques.

WordNet, un lexique sémantique développé à l'université de Princeton par George Miller, offre une alternative intéressante pour dépasser la limite des trois catégories traditionnelles de la NER (Fellbaum 1998). Sur la base d'études expérimentales en psycholinguistique, WordNet reconnaît un certain nombre de principes qui régissent la mémoire lexicale et propose un ensemble de 44 catégories sémantiques. À titre d'exemple, plusieurs expériences examinées par Miller (1990) suggèrent des corrélations cognitives entre les temps de réaction cérébrale et la structure hiérarchique du lexique. Ces corrélations permettraient donc d'inscrire tous les concepts possibles d'une phrase, de façon exhaustive, dans l'ensemble formé par ces 44 dénominateurs sémantiques. L'ambition des concepteurs de WordNet est donc que ces catégories soient transférables à toutes les langues naturelles.

Sur la base de WordNet, Ciaramita et Johnson (2003) ont développé un *SuperSense Tagger* (SST) pour l'anglais. Cette application technologique a été également adoptée pour l'apprentissage automatique d'ontologies par Picca et al. (2008 ; 2009). Ciaramita et Johnson ont appelées « supersenses » les 44 catégories lexicographiques de WordNet. Dans le processus de la NER, les *supersenses* parviennent à établir des liens de proximité sémantique entre des mots de signification proche en en réduisant l'ambiguïté (Ciaramita et Altun 2006 ; Ciaramita et Atserias 2007).

Par exemple, en anglais, le mot *folk* a quatre sens différents mais, comme on le voit dans l'application de WordNet téléchargeable sur le site de l'Université de Princeton, seulement deux *supersenses*, à savoir *noun.group* et *noun.communication* :

1. People in general (noun.group).
2. A social division of (usually preliterate) people (noun.group).
3. People descended (noun.group) from a common ancestor.
4. The traditional and typically anonymous music that is an expression of the life of people in a community (noun.communication).

La disponibilité de *WordNet* et du *SuperSense Tagger* nous a permis de développer un dispositif de marquage automatisé pour l'italien. Nous avons adapté la technologie SST à l'italien en utilisant un algorithme d'apprentissage supervisé (*Supervised Learning*). La base de données d'apprentissage utilisée pour cet algorithme est le corpus MultiSemCor. Il s'agit d'un « set of training examples » composé de 116 textes en italien traduits à partir des textes anglais contenus dans la base de données SemCor (Collins 2002; Bentivogli et al. 2004 ; Nguyen et Guo 2007). Une fois marqué selon les catégories de Word-

net, ce corpus nous a permis d'activer la fonction de prédiction du dispositif. Il est évident que l'efficacité de la classification produite par IPSE est fortement dépendante de la qualité et de la quantité des « training data ». Cependant, pour cette première phase de conception, le set de 116 textes satisfait à nos exigences. La dernière version du WordNet italien contient environ 58000 sens de mots italiens et 41500 lemmes organisés en 32700 *synsets* (groupes synonymiques) alignés avec les *synsets* du WordNet anglais. Les résultats sont presque comparables à ceux obtenus pour l'anglais. Pour le marquage de catégories telles que *noun.person, noun.body* ou *noun.time*, notamment, nous avons pu estimer un taux de précision moyenne supérieur à 70%.

IPSE est actuellement accessible en ligne.

Description de IPSE

Le dispositif proposé est destiné à faciliter la collecte de données simples à partir de documents sélectionnés et, dans la dynamique de l'interprétation individuelle, à en rendre possible la comparaison statistique. En plus de l'habituelle reconnaissance de séquences de caractères de l'alphabet latin encodé selon la norme ASCII (*American Standard Code for Information Interchange*), IPSE permet aussi l'extraction automatisée des catégories générales, sémantiques et logiques, dont on a parlé précédemment (*SuperSenses*). Le but de cette opération est celui de relier et de comparer, sur la base de ces catégories générales, plusieurs textes assemblés dans un corpus librement configuré. Par le marquage automatique de textes en langue italienne, effectué sur la base des 44 catégories définies dans WordNet, on obtient un taux moyen de fiabilité qui se situe autour de 63%. Après cette phase encore expérimentale, nous nous proposons d'élargir l'application à d'autres langues et de permettre, à partir d'une même interface, la comparaison sémantique de textes écrits en langues différentes.

Notre constat de départ à été que les outils d'interrogation électronique normalement utilisés pour la littérature italienne sont des bases de données fermées, limitées en quantité et en qualité des textes. L'outil le plus souvent utilisé dans les études italiennes, pas seulement par les spécialistes, est encore LIZ 4.0 (Stoppelli et Picchi 2001) : un CD contenant 1000 textes de la littérature italienne depuis ses origines aux années 1930 (la limite du droit d'auteur). Les deux principaux inconvénients de ce type d'outils sont les suivants: 1) le nombre limité de textes inclus, pas flexible à l'égard d'autres documents qui pourraient être d'intérêt pour l'utilisateur (par exemple, aujourd'hui, les livres achetés en format numérique pourraient être interrogés à titre privé); 2) la délimitation strictement disciplinaire du

choix des textes. Cette « fermeture » offre, bien évidemment, aussi des avantages, à savoir la garantie du contrôle et de la fiabilité de son contenu.

On a pensé, par conséquent, à la possibilité de créer une base de données polyvalente, qui ne se limite pas à un intérêt disciplinaire spécifique, qui soit progressivement extensible et flexible par rapport aux documents proposés par chaque utilisateur. L'interface et le type de fonctionnalités visent un utilisateur moyen qui soit mis en condition d'adapter l'outil à ses besoins particuliers. L'application n'est pas finalisée à une recherche prédéfinie mais offre à l'utilisateur la possibilité de trouver la pertinence de l'instrument dans l'acquisition d'informations utiles aux objectifs qui lui sont propres.

Documents et corpus

Le premier pas pour l'utilisateur consiste à charger sur le serveur des documents au format TXT et à les réunir en créant un ou plusieurs corpus. Lorsque des documents sont d'intérêt général et suffisamment fiables, l'éditeur du programme a la possibilité de les acquérir en permanence. En ce sens, par exemple, on peut envisager la création d'un corpus de textes de la littérature italienne. Lorsque l'utilisateur charge un document, il lui est demandé de remplir un formulaire où plusieurs informations sur le document doivent être données, y compris la source. Voici un exemple de fiche d'information

Les documents chargés et les corpus créés par les utilisateurs sont, par la suite, disponibles sous forme de liste. Comme le montre l'illustration ci-dessus, les documents restent toujours accessibles au format « Fichier texte » et au format « Fichier tags ». Grâce à la version « fichier tags », il est donc possible de vérifier en détail la NER réalisée automatiquement par le dispositif.

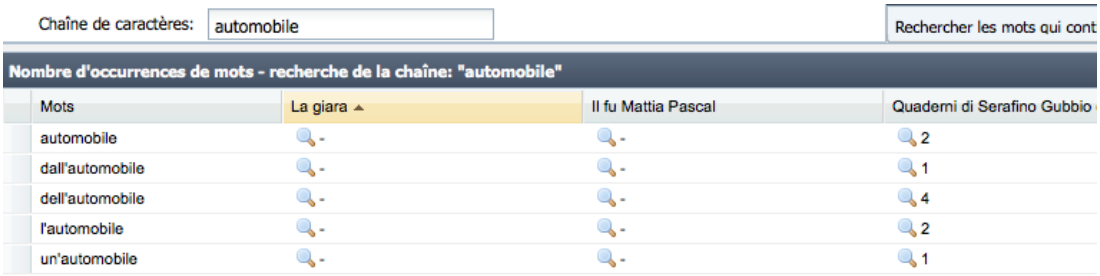
Type de recherches

Les recherches qu'on peut effectuer sur un corpus sont de trois types: par chaîne de caractères, par fonctions grammaticales, par catégories sémantiques.

1) *Recherche par chaîne de caractères*. Il s'agit d'une recherche de type traditionnel, effectuée pour un seul mot ou pour une partie de mot en utilisant l'astérisque et le point d'interrogation à la place d'un ou plusieurs éléments variables. La requête montrera exactement combien de fois la séquence ou le mot apparaissent dans les textes et dans le corpus. Avant d'effectuer la recherche, on peut spécifier un « seuil significatif », qui est le nombre minimum d'occurrences jugées utiles. Les occurrences trouvées peuvent être vérifiées dans le contexte du document en cliquant sur l'icone d'une petite loupe. Par exemple, pour le corpus « Pirandello prosa », échantillon composé de trois œuvres de l'auteur, la recherche de la séquence « tram » permet de vérifier la présence linguistique de ce moyen de transport. En indiquant comme seuil de pertinence la valeur 1, on obtient, grâce à la fonction « Exporter vers Excel », le tableau suivant:

Occurences of words						
	La giara		Il fu Mattia Pascal		Quaderni di Serafino Gubbio operatore	
entrambi	25.00%	2	18.75%	3	6.25%	1
tram	12.50%	1	18.75%	3	12.50%	2
trambasciata	12.50%	1	0.00%	0	0.00%	0
tramenio	12.50%	1	0.00%	0	0.00%	0
tramontata	12.50%	1	0.00%	0	0.00%	0
tramonti	12.50%	1	0.00%	0	0.00%	0
tramonto	12.50%	1	6.25%	1	0.00%	0
concentrammo	0.00%	0	6.25%	1	0.00%	0
d'ammaestramento	0.00%	0	6.25%	1	0.00%	0
d'entrambi	0.00%	0	6.25%	1	0.00%	0
entrambe	0.00%	0	12.50%	2	6.25%	1
sinistramente	0.00%	0	0.00%	0	12.50%	2
stramazza	0.00%	0	0.00%	0	12.50%	2
stramazzт	0.00%	0	0.00%	0	6.25%	1
stramberie	0.00%	0	6.25%	1	0.00%	0
strambo	0.00%	0	6.25%	1	6.25%	1
strampalate	0.00%	0	6.25%	1	6.25%	1
strampalati	0.00%	0	6.25%	1	0.00%	0
trama	0.00%	0	0.00%	0	6.25%	1
tramenмo	0.00%	0	0.00%	0	6.25%	1
tramestмo	0.00%	0	0.00%	0	6.25%	1
tramutata	0.00%	0	0.00%	0	6.25%	1
tramviarie	0.00%	0	0.00%	0	6.25%	1

En plus de la présence du mot « tram », on peut également observer dans le bas du tableau la présence de l'adjectif « tramviarie ». Dans une recherche de ce type, les pourcentages sont évidemment peu significatifs. Si, par exemple, on compare ce résultat avec la présence du terme «automobile», on note que, à la différence du cas précédent, le terme n'apparaît que dans le troisième livre. On peut le voir ci-dessous dans l'écran d'utilisation qui précède l'éventuelle exportation vers Excel :



2) *Recherche par fonctions grammaticales*. Wordnet classe les parties du discours en huit groupes: adjectifs, conjonctions, éléments qui n'appartiennent pas à d'autres groupes (articles, adverbes, interjections, symboles et chiffres)[2], noms, ponctuation, prépositions, pronoms, verbes. Pour la catégorie «adjectifs», par exemple, les trois valeurs indiquées correspondent au pourcentage d'éléments reconnus pour chaque sous-catégorie a) dans un seul document, b) dans le corpus analysé ou c) comme total arithmétique des occurrences.

[2] Cette typologie est basée sur l'anglais, les articles sont donc regroupés avec d'autres éléments invariants. Cet aspect n'a pas d'incidence sur l'application à l'italien.

Fonctions Grammaticales	La giara			Il fu Mattia Pascal			Quaderni di Serafino Gubbio operato		
adjective									
qualifying	5.517 %	28.704 %	3625	5.388 %	38.744 %	4893	5.634 %	32.552 %	4111
numeral	1.230 %	37.988 %	808	0.816 %	34.838 %	741	0.792 %	27.174 %	578
possessive	1.164 %	22.966 %	765	1.651 %	45.002 %	1499	1.462 %	32.032 %	1067
indefinite	1.064 %	28.129 %	699	1.021 %	37.304 %	927	1.177 %	34.567 %	859
desmonstrative	0.892 %	24.225 %	586	1.057 %	39.686 %	960	1.196 %	36.089 %	873
interrogative or exclamative	0.070 %	27.545 %	46	0.076 %	41.317 %	69	0.071 %	31.138 %	52

La ligne des adjectifs montre que la lemmatisation des formes fléchies (singulier, pluriel, masculin, féminin) n'est pas encore automatisée de manière complètement efficace. On voit qu'elle fonctionne pour la forme en tête de la liste, « vecchio », et ne l'est pas pour « nuovo », le deuxième élément par ordre de fréquence. Cependant, pour les formes dont la flexion est limitée, comme les adjectifs et les substantifs, les résultats sont utilisables à partir d'un rapide croisement des données fournis par le programme[3].

Fréquences d'apparition des lemmes - Fonction grammaticale: "adjective - qualifying"

Lemmes	La giara	Il fu Mattia Pascal	Quaderni di Serafino Gubbio operato
vecchio	2.905 % (50)	1.3232 % (32)	0.8017 % (17)
	La giara: vecchi vecchie vecchio vecchio vecchio		
	Il fu Mattia Pascal: vecchi vecchie vecchio vecchio vecchio		
	Quaderni di Serafino Gubbio operato: vecchi vecchie vecchio		
nuovo	2.6145 % (45)	1.5337 % (37)	0.8518 % (18)
	La giara: nuovo nuovo		
	Il fu Mattia Pascal: nuovo nuovo		
	Quaderni di Serafino Gubbio operato: nuovo nuovo		

3) *Recherche par catégories sémantiques*. Il s'agit de la fonction de recherche la plus étendue du dispositif. Wordnet, conçu au départ comme un réseau de liens sémantiques (*Dictionary Browser*), permet, grâce à l'activation de ses 44 dénominateurs sémantiques, de produire un dispositif pertinemment applicable à des documents linguistiquement complexes. Les deux principales macro-catégories développées par Wordnet sont les « noms » et les « verbes ». Les sous-catégories marquées automatiquement sont 26 pour les noms et 15 pour les verbes, comme indiqué dans le tableau qui résume la totalité des dénominateurs sémantiques.

[3] Pour les verbes, à ce stade, la lemmatisation reste encore problématique.

	Name	**Contents**
0	adj.all	all adjective clusters
01	adj.pert	relational adjectives (pertainyms)
02	adv.all	all adverbs
03	noun.Tops	unique beginner for nouns
04	noun.act	nouns denoting acts or actions
05	noun.animal	nouns denoting animals
06	noun.artifact	nouns denoting man-made objects
07	noun.attribute	nouns denoting attributes of people and objects
08	noun.body	nouns denoting body parts
09	noun.cognition	nouns denoting cognitive processes and contents
10	noun.communication	nouns denoting communicative processes and contents
11	noun.event	nouns denoting natural events
12	noun.feeling	nouns denoting feelings and emotions
13	noun.food	nouns denoting foods and drinks
14	noun.group	nouns denoting groupings of people or objects
15	noun.location	nouns denoting spatial position
16	noun.motive	nouns denoting goals
17	noun.object	nouns denoting natural objects (not man-made)
18	noun.person	nouns denoting people
19	noun.phenomenon	nouns denoting natural phenomena
20	noun.plant	nouns denoting plants
21	noun.possession	nouns denoting possession and transfer of possession
22	noun.process	nouns denoting natural processes
23	noun.quantity	nouns denoting quantities and units of measure
24	noun.relation	nouns denoting relations between people or things or ideas
25	noun.shape	nouns denoting two and three dimensional shapes
26	noun.state	nouns denoting stable states of affairs
27	noun.substance	nouns denoting substances
28	noun.time	nouns denoting time and temporal relations
29	verb.body	verbs of grooming, dressing and bodily care
30	verb.change	verbs of size, temperature change, intensifying, etc.
31	verb.cognition	verbs of thinking, judging, analyzing, doubting

32	verb.communication	verbs of telling, asking, ordering, singing
33	verb.competition	verbs of fighting, athletic activities
34	verb.consumption	verbs of eating and drinking
35	verb.contact	verbs of touching, hitting, tying, digging
36	verb.creation	verbs of sewing, baking, painting, performing
37	verb.emotion	verbs of feeling
38	verb.motion	verbs of walking, flying, swimming
39	verb.perception	verbs of seeing, hearing, feeling
40	verb.possession	verbs of buying, selling, owning
41	verb.social	verbs of political and social activities and events
42	verb.stative	verbs of being, having, spatial relations
43	verb.weather	verbs of raining, snowing, thawing, thundering
44	adj.ppl	participial adjectives

Comme on le voit dans le tableau ci-dessous, pour les dix premiers résultats de la catégorie « sentiments » (« feeling »), par exemple, le marquage automatique est très fiable et toutes les occurrences sont pertinentes. Dans d'autres cas, la précision est inférieure et des résultats non pertinents se trouvent parmi les occurrences à fréquence élevée. En attendant que le fichier de reconnaissance linguistique soit perfectionné, l'utilisateur peut vérifier et utiliser les catégories le plus efficaces.

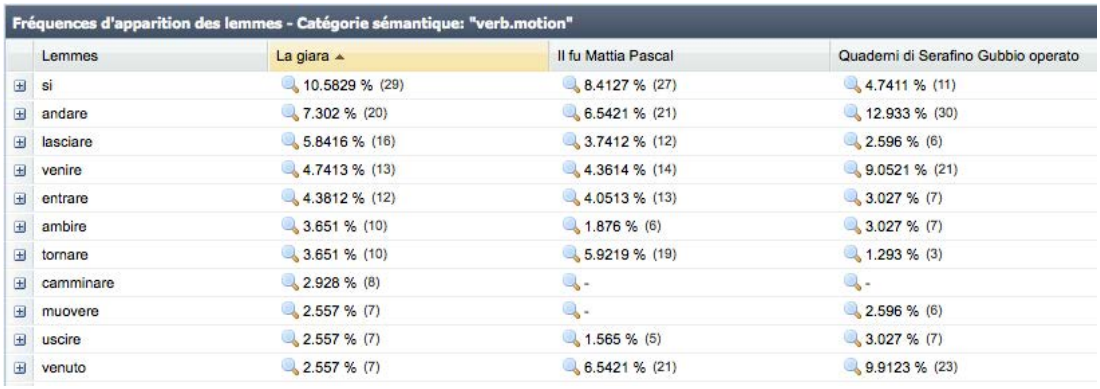
Fréquences d'apparition des lemmes - Catégorie sémantique: "noun.feeling"

Lemmes	La giara ▲	Il fu Mattia Pascal	Quaderni di Serafino Gubbio operato
piacere	15.381 % (10)	9.3812 % (12)	15.1323 % (23)
rabbia	15.381 % (10)	8.5911 % (11)	9.2114 % (14)
paura	13.859 % (9)	16.4121 % (21)	9.8715 % (15)
amore	10.777 % (7)	7.039 % (9)	7.2411 % (11)
gusto	9.236 % (6)	4.696 % (6)	2.634 % (4)
speranza	7.695 % (5)	6.258 % (8)	1.973 % (3)
vergogna	7.695 % (5)	-	5.268 % (8)
terrore	6.154 % (4)	-	6.581 % (10)
disprezzo	4.623 % (3)	-	4.617 % (7)
gratitudine	4.623 % (3)	2.343 % (3)	-

Pour les verbes de mouvement (« motion »), le marquage automatique est aussi satisfaisant. On peut en observer les premiers résultats dans la liste :

À la première ligne, le résultat « si » n'est clairement pas approprié, ce qui s'explique probablement par la position du mot, qui précède souvent une forme verbale. Les dix occurrences qui suivent, par contre, indiquent toutes des verbes de mouvement. Pour ces formes la lemmatisation automatique fonctionne avec quelques exceptions : pour le verbe « venir », par

exemple, le singulier du participe passé n'est pas intégré dans la ligne de la forme à l'infinitif.



Fréquences d'apparition des lemmes - Catégorie sémantique: "verb.motion"

Lemmes	La giara	Il fu Mattia Pascal	Quaderni di Serafino Gubbio operato
si	10.5829 % (29)	8.4127 % (27)	4.7411 % (11)
andare	7.302 % (20)	6.5421 % (21)	12.933 % (30)
lasciare	5.8416 % (16)	3.7412 % (12)	2.596 % (6)
venire	4.7413 % (13)	4.3614 % (14)	9.0521 % (21)
entrare	4.3812 % (12)	4.0513 % (13)	3.027 % (7)
ambire	3.651 % (10)	1.876 % (6)	3.027 % (7)
tornare	3.651 % (10)	5.9219 % (19)	1.293 % (3)
camminare	2.928 % (8)	-	-
muovere	2.557 % (7)	-	2.596 % (6)
uscire	2.557 % (7)	1.565 % (5)	3.027 % (7)
venuto	2.557 % (7)	6.5421 % (21)	9.9123 % (23)

Un exemple d'interrogation interprétative: Arioste pétrarquiste

À partir d'un échantillon de quatre textes représentatifs, on peut essayer de vérifier le degré de proximité des poèmes de l'Arioste (1474-1533) au modèle de pétrarquisme fixé par Pietro Bembo (1470-1547). Les textes choisis, en plus des *Rime* de l'Arioste, sont l'*Orlando Furioso*, le *Canzoniere* de Pétrarque (1304-1374) et les *Rime* de Bembo. Par cet exemple, on peut remarquer que le dispositif fonctionne assez bien aussi pour l'italien ancien. Un élément caractéristique du pétrarquisme est la représentation idéalisée de la femme. La dénomination des parties du corps féminin, extrêmement codifiée dans les pratiques traditionnelles de la poésie d'amour courtois, peut être comparée à travers l'interrogation sémantique. Dans la partie haute de l'écran produit par la requête, IPSE fournit le tableau suivant :

Les deux premiers éléments, « occhi » et « viso », typiques du pétrarquisme, montrent un équilibre certain dans les pourcentages d'utilisation propres aux trois recueils de poèmes d'amour. Ce n'est pas surprenant que le pourcentage soit plus faible dans le *Roland furieux*, où la variation thématique et lexicale est beaucoup plus large. On remarque un premier écart chez Arioste pour le mot « petto », nettement moins fréquent en pourcentage chez Pétrarque et chez Bembo. L'absence de « faccia », mot exclu de la langue de la poésie d'amour suite à la codification de Pétrarque, est homogène dans les trois recueils. Arioste, toutefois, utilise de manière sensiblement plus fréquente les mots « collo », « corpo » et « bocca », non représentatifs chez Pétrarque et chez Bembo. Il faut préciser que le terme « bocca » est également présent chez ces deux auteurs, mais l'écran indique la valeur zéro parce que le seuil significatif a été fixé à un minimum de trois occurrences.

Ces quelques constats, bien évidemment, ne montrent qu'une simple tendance, mais suggèrent déjà, pour Arioste, une représentation physique de la femme plus réaliste vis-à-vis du canon de son temps. Ce choix de représentation reste à démontrer à un autre niveau d'analyse, mais l'exemple de recherche, malgré le fait qu'il soit très sommaire, est suffisant ici pour montrer la pertinence des catégories sémantiques appliquées de manière entièrement automatisée.

À une autre échelle, IPSE pourra offrir la possibilité de cartographier et de comparer sémantiquement des corpus de plus vastes dimensions. L'opération est possible à partir d'un filtrage mécanique des informations et donc dans une perspective renversée par rapport à celle de l'individu qui entraîne ses compétences sur des documents imprimés. Pour cette raison, la sélection préalable et objective d'informations fournie par le dispositif demande une profonde humilité dans l'interprétation des données, surtout lorsque l'utilisateur se donne le droit de se référer à un corpus-objet dont il possède une maîtrise cognitive très limitée. Sa stratégie de lecture, avec l'énonciation de ses objectifs, devient alors essentielle pour définir et reconnaître l'objet à analyser.

Bibliographie

BENTIVOGLI Luisa, FORNER Pamela et PIANTA Emanuele (2004), « Evaluating crosslanguage annotation transfer in the multisemcor corpus »*, Proceedings of the 20th international conference on Computational Linguistics (COLING '04),* Association for Computational Linguistics, Stroudsburg, PA, USA, Article 364.

CIARAMITA Massimiliano et ALTUN Yasemin (2006), « Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger », *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP '06),* Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 594-602.

CIARAMITA Massimiliano et ATSERIAS J. (2007), « Pos-tagging with a named entity tagger », *Intelligenza Artificiale*, 4, pp. 28–29.

CIARAMITA Massimiliano et JOHNSON Mark, (2003), « Supersense tagging of unknown nouns in WordNet », *Proceedings of the 2003 conference on Empirical methods in natural language processing (EMNLP '03),* Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 168-175.

COLLINS Michael (2002), « Discriminative training methods for hidden Markov models: theory and experiments with perceptron algorithms », *Proceedings of the ACL-*

02 conference on Empirical methods in natural language processing (EMNLP '02), Vol. 10. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1-8.

FELLBAUM Christiane (1998), *Wordnet. An Electronic Lexical Database*, Cambridge-London, The Massachusetts Institute of Technology Press.

MILLER George A. (1990), « Nouns in Wordnet: a lexical inheritance system », *International Journal of Lexicography*, 3(4), pp. 245–264.

NGUYEN Nam et GUO Yunsong (2007), « Comparisons of sequence labeling algorithms and extensions », in *Proceedings of the 24th international conference on Machine learning (ICML '07)*, éd. Zoubin Ghahramani, ACM, New York, NY, USA, pp. 681-688.

PICCA Davide, GLIOZZO Alfio et CIARAMITA Massimiliano (2008), « SuperSense Tagger for Italian », *Proceedings of the International Conference on Language Resources and Evaluation (LREC '08),* European Language Resources Association, Marrakech, Morocco*,* pp. 2386-2390.

PICCA Davide, GLIOZZO Alfio Massimiliano et CAMPORA Simone (2009), « Bridging languages by SuperSense entity tagging », I *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS '09),* Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 136-142.

SEKINE Satoshi (2004), « Named Entity: History and Future ».

STOPPELLI Pasquale et PICCHI Eugenio (2001), LIZ 4.0 *Letteratura italiana* Zanichelli, CD-ROM, Bologna, Zanichelli.