AIWL: una lista di frequenza dell'italiano accademico

Stefania Spina

Dipartimento di Scienze del linguaggio, Università per stranieri di Perugia

Riassunto

In questo articolo è descritta l'AIWL (*Academic Italian Word List*), una lista di frequenza delle parole italiane non tecniche più usate nella comunicazione accademica scritta. La lista è legata all'esigenza di rafforzare ed espandere il lessico accademico degli studenti di lingua madre non italiana delle Università italiane. L'AIWL è stata estratta da un corpus di italiano scritto accademico; il corpus, che misura 1 milione di parole ed è composto da 240 testi appartenenti ad aree disciplinari e a tipologie testuali diverse, è bilanciato, annotato per categoria grammaticale e lemmatizzato. Le unità lessicali estratte dal corpus (singole parole e combinazioni di parole) sono ordinate per frequenza e selezionate sulla base di un coefficiente statistico di dispersione all'interno delle diverse aree disciplinari rappresentate nel corpus. Lo scopo dell'AIWL è quello di fornire uno strumento computazionale e lessicografico per la costituzione di applicazioni di *natural language processing* da utilizzare in un ambiente di apprendimento di rete. L'articolo descrive i presupposti teorici e le metodologie di estrazione della lista di frequenza, cercando di descrivere alcune caratteristiche delle unità lessicali presenti nell'AIWL.

Abstract

In this paper I describe the *Academic Italian Word List* (AIWL), a frequency list of the most common non-technical words used in written academic communication. The project arises from the need to expand academic vocabulary of non-native students of Italian Universities. The AIWL is a corpus-based list, being extracted from a balanced, POS-tagged and lemmatized corpus of Italian academic written language (the AIC, *Academic Italian Corpus*). The AIC includes 1 million words and is composed of 240 texts belonging to different subject areas and textual typologies. The lexical units extracted from the AIC (single words as well as word combinations) are ordered by frequency and selected by a statistical measure of dispersion within the different subject areas. The AIWL aims to provide a computational and lexicographical resource to support the constitution of natural language processing applications to be used in an online learning environment. This paper describes in detail the theoretical assumptions and the methodology of extraction of the frequency list, and it outlines the main features of the lexical units that are included in the AIWL.

Keywords: academic vocabulary, frequency list, dispersion, academic corpus, collocations

1. Introduzione

Questo studio trae la sua origine dalla necessità di misurare e successivamente ampliare la competenza del lessico accademico italiano da parte di studenti universitari di madrelingua diversa dall'italiano.

Per frequentare corsi universitari in Italia gli studenti con lingua madre diversa dall'italiano devono superare dei test che ne certifichino il livello di competenza linguistica. Il livello minimo richiesto è il B2 del *Common European Framework* (Council of Europe, 2001).

Studi precedenti, tuttavia (Spina, forthcoming,b; Iengo, 2009), hanno dimostrato che una percentuale consistente di studenti non nativi hanno una conoscenza molto imprecisa del lessico

accademico italiano, sia a livello produttivo sia ricettivo. Conseguenze principali di questa competenza imperfetta sono i significativi problemi di comprensione e produzione evidenziati dagli studenti nel corso delle attività accademiche; tali problemi riguardano sia studenti con L1 distanti dall'italiano (test lessicali effettuati su studenti cinesi hanno mostrato ad esempio che sono in grado di comprendere circa un terzo del lessico accademico italiano; Jie, 2009), sia quelli europei o addirittura di lingue madri romanze.

2. Obiettivi

Date queste premesse, l'Università per Stranieri di Perugia ha dato avvio a un progetto di ricerca per identificare, con metodologie basate sull'uso di corpora e di strumenti statistici, il lessico accademico italiano. L'obiettivo primario è quello di creare una lista di frequenza accademica (*Academic Italian Word List*, o AIWL), a partire da un corpus scritto di lingua italiana accademica. La lista di frequenza, come altre omologhe già costituite per altre lingue (Coxhead, 2000; Paquot, 2007, ad esempio, per l'inglese), è mirata ad identificare le parole che ricorrono comunemente in un vasto insieme di testi accademici ma che al contrario non sono così frequenti in testi non accademici. L'AIWL include i 403 lemmi e le 208 collocazioni più frequenti nel lessico accademico italiano scritto.

Oltre a poter essere utilizzata per lo sviluppo di materiali per la didattica e la valutazione dell'italiano accademico, la lista di frequenza servirà come base per lo sviluppo di applicazioni di *natural language processing* per l'addestramento e il potenziamento di questa tipologia di lessico all'interno di un ambiente di apprendimento di rete; in particolare, sulla base della lista di frequenza accademica è stato costituito un *database* lessicale che sarà integrato in un *Personal Learning Environment*, consentendo di fornire supporto alle attività di comprensione e di produzione degli studenti e di automatizzare alcune operazioni di *testing* della competenza del lessico accademico (Spina, forthcoming,a).

3. Metodologia

3.1. Lessico accademico

La definizione di lessico accademico è basata sugli studi di Nation, che sostiene che il vocabolario di una lingua può essere suddiviso in quattro sezioni distinte (Nation, 2001):

1. Lessico di alta frequenza: per l'italiano, il *Vocabolario di base* (De Mauro, 1980).
2. Lessico accademico: parole con frequenza elevata in tutti i testi accademici, a prescindere dalla disciplina.
3. Lessico tecnico: parole specifiche di singole aree tematiche e discipline.
4. Lessico di bassa frequenza.

Il lessico accademico, dunque, è strettamente connesso con le attività didattiche e di apprendimento che gli studenti svolgono comunemente in ambito universitario (seguire lezioni e seminari, studiare manuali, dispense, appunti, articoli scientifici, sostenere esami scritti e orali); la corretta interpretazione di unità lessicali come *approccio*, *contesto*, *articolato*, *eterogeneo*, *metodologia* o *introdurre un concetto* è indispensabile per poter portare a termine con successo tali attività.

La competenza del lessico accademico, inoltre, deve necessariamente coprire sia il livello ricettivo sia quello produttivo, perché in entrambe queste modalità comunicative esso è utilizzato dagli studenti universitari.

3.2. Academic Italian Corpus [1]

L'AIWL si rifà, almeno nelle premesse, al modello dell'AWL (*Academic Word List*), la lista di frequenza del lessico accademico inglese (Coxhead, 2000), che contiene le 570 famiglie di parole (Bauer and Nation, 1993) che non sono comprese nelle 2.000 più frequenti della *General Service List* (West, 1953).

Come quella inglese, l'estrazione della lista di frequenza italiana è stata portata a termine in un corpus scritto di testi accademici (l'*Academic Italian Corpus*, o AIC). Il corpus, che misura 1 milione di parole, è stato costituito attraverso due tipi diversi di bilanciamento dei testi:

- un bilanciamento orizzontale (le aree tematiche di appartenenza dei testi);
- un bilanciamento verticale (le differenti tipologie testuali accademiche).

La composizione del corpus è schematizzata in Tab. 1.

Per quanto riguarda le aree tematiche di appartenenza dei testi, ne sono state individuate tre, a loro volta suddivise in sotto-discipline più specifiche:

- area umanistica (storia, letteratura, linguistica, arte ecc.);
- area giuridico-economica (diritto, economia, scienze politiche ecc.);
- area scientifica (medicina, informatica, matematica ecc.).

Ciascuna delle tre aree comprende 80 testi (bilanciati nelle varie sotto-discipline) di poco più di 4000 parole ciascuno.

I testi che compongono il corpus sono dunque in totale 240 ed appartengono, secondo il bilanciamento verticale previsto dal progetto, a 4 generi diversi, tutti collegati con l'ambito accademico: manuali, dispense, articoli scientifici e tesi di laurea.

	Manuali	*Dispense*	*Articoli*	*Tesi di laurea*	*Totale*
Area umanistica (333.580 parole)	10	35	20	15	80
Area giuridico-economica (333.970 parole)	10	35	20	15	80
Area scientifica (333.120 parole)	10	35	20	15	80
Totale Testi	*30*	*105*	*60*	*45*	*240*

Tabella 1: Composizione del corpus AIC

Il corpus raccolto sulla base di tale metodologia è stato successivamente annotato per categoria grammaticale e lemmatizzato; per le operazioni di tokenizzazione, POS-tagging e lemmatizzazione è stato utilizzato *TreeTagger* (Schmid, 1994), con un *tagset* e un *file* di parametri per l'italiano elaborati all'Università per Stranieri di Perugia [2].

Una volta etichettato per categoria grammaticale e lemmatizzato, il corpus è stato successivamente trattato per poter essere interrogato attraverso l'IMS *Corpus Workbench* (Christ, 1994), la *suite* di strumenti per l'analisi automatica di corpora linguistici elaborata a Stoccarda e basata sul linguaggio di interrogazione CQP (*Corpus Query Processing*) [3].

[1] Il sito del progetto è http://elearning.unistrapg.it/corpora/.

[2] Il file di parametri, modellato su quello di Marco Baroni (ftp://ftp.ims.uni-stuttgart.de/pub/corpora/italian-par2-linux-3.1.bin.gz), prevede un *tagset* con 50 etichette di categorie grammaticali, ed è anche usato nella versione web di *TreeTagger* sviluppata da Francesco Scolastra a Perugia (http://elearning.unistrapg.it/TreeTagger.html).

[3] *IMS Corpus Workbench* è descritto in dettaglio anche nel sito Internet http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/.

3.3. La lista di frequenza dell'italiano accademico (AIWL)

Attraverso un'interrogazione dell'AIC è stata estratta una lista dei 19.890 lemmi che compongono il corpus, ordinata per frequenza decrescente. Tale lista iniziale è stata in seguito ristretta ai soli 6.175 lemmi presenti in tutte e tre le aree tematiche del corpus, e successivamente ai circa 5.200 che abbiano frequenza ≥ 10.

La distribuzione in tutte le aree tematiche che costituiscono il corpus è fondamentale per escludere parole tecniche, tipiche solo di specifiche discipline: così, ad esempio, il lemma *membrana*, che ha una frequenza elevata nella sezione delle scienze (115), è stato filtrato nella lista che tiene conto della distribuzione, perché assente nelle sezioni umanistica e giuridico-economica.

La distribuzione in tutte le aree tematiche, tuttavia, non risolve completamente il problema della sovra-rappresentazione in una delle tre e della presenza solo occasionale nelle altre due. Prendiamo come esempio i casi dei lemmi *sintattico* e *omogeneo*, che hanno nel corpus accademico la seguente distribuzione (Tab. 2):

	Area umanistica	*Area giuridico-economica*	*Area scientifica*
sintattico	59	2	11
omogeneo	22	14	33

Tabella 2: Distribuzione dei lemmi sintattico *e* omogeneo *nel corpus AIC*

È evidente che *omogeneo* ha una distribuzione molto più uniforme di *sintattico*, che presenta una frequenza elevata nella sezione umanistica a causa della sua alta frequenza nei testi di linguistica, ma frequenze più basse o addirittura occasionali nelle altre due sezioni.

Per selezionare solo i lemmi con una distribuzione uniforme si è scelto di utilizzare il coefficiente di dispersione D elaborato da Juilland e Chang Rodriguez (Bortolini et al., 1971; Oakes, 1998); i lemmi con $D \leq 0{,}7$ sono stati esclusi dalla lista di frequenza. Nell'esempio precedente, di conseguenza, *omogeneo* è stato inserito mentre *sintattico* non ha raggiunto il valore minimo di dispersione ed è stato escluso.

La lista risultante da questi tre processi di selezione (presenza in tutte le tre aree tematiche, frequenza ≥ 10 e $D \geq 0{,}7$) è stata denominata *pre-AIWL*; il passo successivo è stato il confronto con l'elenco integrale delle parole che fanno parte del *Vocabolario di base* (De Mauro, 1980): le 6.522 parole, suddivise tra fondamentali, di alto uso e di alta disponibilità, più frequenti dell'italiano.

Il risultato di questa operazione di *matching* sono i 403 lemmi presenti nella lista accademica e assenti nel *Vdb*, cioè la lista di frequenza dell'italiano accademico [4]; si tratta dei lemmi più uniformemente utilizzati nella comunicazione accademica scritta, ulteriori rispetto al vocabolario di base e quindi più caratterizzanti il lessico universitario nel suo insieme, a prescindere dal settore disciplinare. Il fatto che la lista accademica italiana contenga un numero di parole più ridotto rispetto a quella inglese (che comprende 571 lemmi) può essere spiegato dal fatto che

[4] Tra i lemmi presenti sia in *Vdb* che in *pre-AIWL* e, quindi, non inseriti in AIWL, ci sono ovviamente alcune delle parole più frequenti dell'italiano, con alta frequenza anche nel lessico accademico (*sistema*, *forma*, *esempio*, *trattare*, *diverso*, *generale* ecc.). Vi sono tuttavia anche circa 1600 lemmi presenti nel *Vdb* ma assenti in *pre-AIWL*, quindi assenti, poco frequenti o senza una distribuzione uniforme nel corpus accademico (ad esempio *zitto*, *valigia*, *abbagliare*, *succulento*, *supplicare* ecc.).

il *Vdb* contiene 6.522 parole, contro le sole 2.000 della *General Service List* inglese. L'italiano etichetta dunque un numero di alto di parole come "di base" rispetto all'inglese.

3.4. Collocazioni accademiche

È stato osservato (Paquot, 2007) come le liste di frequenza accademiche esistenti siano più orientate alla ricezione che alla produzione, anche perché comprendono solo elementi lessicali formati da singole parole; è il caso, ad esempio, dell'AWL (Coxhead, 2000). Questa scelta è in contraddizione con le tesi più accreditate della linguistica acquisizionale, che evidenziano l'importanza della memorizzazione e dell'uso attivo di elementi lessicali "prefabbricati", composti da più di una parola, sia per i parlanti nativi sia per i non nativi. È stato dimostrato che la "competenza collocazionale" svolge un ruolo centrale:

- nel migliorare la fluenza dei parlanti non nativi;
- nel facilitare la loro comprensione (Nesselhauf, 2005).

Studi piuttosto recenti sull'inglese hanno inoltre evidenziato che esiste una fraseologia specifica dello scritto accademico (Biber, 2004; Oakey, 2002; Cowie, 1997), collegata in particolar modo ad alcune delle funzioni testuali e informative più utilizzate in ambito accademico ("esemplificare", "fare ipotesi", "trarre conclusioni" ecc.).

In quest'ottica, e in quella parallela di orientare maggiormente l'AIWL in senso non esclusivamente ricettivo, rendendola uno strumento adatto a descrivere il lessico produttivo dell'italiano accademico (Vongpumivitch et al., 2009), dal corpus AIC è stata estratta anche una lista delle collocazioni più frequenti [5].

Sono state prese in considerazione al momento solo le combinazioni di parole che veicolano un contenuto referenziale ("*referential phrasemes*", secondo la classificazione proposta da Granger and Paquot, 2008), escludendo quindi tutte le espressioni con funzione connettiva, testuale o pragmatica.

Alle 403 parole singole si aggiungono quindi 208 collocazioni estratte per cinque delle sequenze di categorie grammaticali più produttive a livello di combinazioni lessicali. La metodologia impiegata è la stessa di quella che è alla base del *Dictionary of Italian Collocations* (Spina, 2010): da un'etichettatura manuale di una lista di collocazioni estratte da corpora dell'italiano contemporaneo sono state individuate le sequenze di categorie grammaticali più frequenti, e, per il lessico accademico, al loro interno sono state selezionate le prime cinque:

1. aggettivo-nome (*netta distinzione*);
2. nome-aggettivo (*prospettiva teorica*);
3. nome-preposizione-nome (*bagaglio di conoscenze*);
4. verbo-articolo-nome (*affrontare il tema*);
5. verbo-nome (*fare riferimento*).

Le collocazioni che rientrano in queste cinque sequenze sono state filtrate, come nel caso delle parole singole, in base al loro coefficiente di dispersione D, sulla base del quale è stato successivamente calcolato il loro coefficiente d'uso (Bortolini et al., 1971). È stata scelta come soglia

[5] La diversificazione terminologica è molto ampia riguardo alle unità lessicali composte da più di una parola; in questo articolo il termine *collocazione* è usato prevalentemente come iperonimo di tutte le tipologie diverse di tali combinazioni, e non come iponimo indicante un tipo particolare di espressioni, le *collocazioni ristrette* (Granger and Meunier, 2008).

di immissione nel *database* del lessico accademico, in conformità con quanto è avvenuto per altri lessici di frequenza dell'italiano scritto come il *LIF* (Bortolini et al., 1971), un coefficiente d'uso > 2, che porta appunto al numero totale di 208 collocazioni.

A partire dall'AIWL (lemmi singoli e collocazioni), è stato successivamente costituito un *database* lessicale (Tab. 3) in cui, ad ogni lemma accademico, sono state associate informazioni di tipo statistico-quantitativo (frequenza nelle singole sezioni del corpus e frequenza globale, indice di dispersione), lessico-semantico (definizione ed esempi d'uso tratti dal corpus) e sintattico (composizione interna delle collocazioni, informazioni sulla loro variabilità/invariabilità, contesto sintattico con cui ciascun lemma co-occorre più frequentemente. In Tab. 3, ad esempio, è contenuta l'informazione che la collocazione *dare luogo* è composta da un verbo e da un nome, che possono essere interrotti dalla presenza di un avverbio, come in *danno spesso luogo*).

Lemmi	*Frequenze e dispersione*					*Grammatica e sintassi*			*Semantica*	
	Sc. Um.	*Sc.*	*Sc. Giu. Eco.*	*Tot.*	*D*	*POS*	*Forme*	*struttura interna*	*Def.*	*Esempio*
approccio	80	90	151	321	0,79	nome	approccio approcci		[def.]	[esempio]
dare luogo	16	22	16	54	0,88	verbo	[elenco forme]	V [ADV] N	[def.]	[esempio]

Tabella 3: Esempio di entrate del database lessicale

4. Analisi

4.1. Singole parole

I lemmi del lessico accademico vanno dunque a ritagliare, all'interno della fascia di lessico comprendente i 47.000 lemmi definiti "comuni" (De Mauro, 1999), una sezione che caratterizza in modo specifico la comunicazione scritta accademica [6]. L'AIWL copre circa il 5% di un testo accademico scritto (e solo lo 0,5% di un testo parlato di registro colloquiale), mentre dello stesso testo accademico il *Vdb* copre circa il 78%. Lessico di base e lessico accademico insieme coprono dunque circa l'83% del vocabolario usato nella comunicazione accademica; il restante 17% è formato dai tecnicismi, che sono ovviamente frequenti e significativi in quest'ambito, e da parole di bassa frequenza.

I 403 lemmi dell'AIWL sono suddivisi, in base alla loro categoria grammaticale, come descritto in Tab. 4.

Nomi	191
Aggettivi	134
Avverbi	45
Verbi	32
Preposizioni	1
Totale	*403*

Tabella 4: Composizione grammaticale dell'AIWL

[6] Resta del tutto aperto, come per altri lessici di frequenza, il problema dei differenti sensi di parola che i lemmi accademici possono assumere a seconda dei contesti d'uso. A questo tema sarà dedicato un segmento successivo del progetto di ricerca sul lessico accademico.

I nomi sono ovviamente la categoria grammaticale più rappresentata; si tratta per lo più di nomi astratti (*approccio*, *coerenza*, *dottrina*, *identità*), spesso derivati (*acquisizione*, *apprendimento*, *approfondimento*, *suddivisione*). Alcuni si riferiscono a livello metalinguistico alle parti di un testo (*capitolo*, *grafico*, *paragrafo*), altri a elementi o attività caratteristiche della vita accademica (*curriculum*, *test*, *insegnamento*).

Tra gli aggettivi, piuttosto numerosi, che molto spesso descrivono proprietà astratte dei nomi a cui si riferiscono (*coerente*, *strutturato*, *sintetico*) dominano quelli in *–ivo* (*cognitivo*, *oggettivo*), in *–ico* (*didattico*, *dinamico*) e in *–ale* (*potenziale*, *sostanziale*).

I verbi sono invece presenti in numero limitato nella lista accademica e rappresentano spesso la variante più elevatà rispetto a forme verbali di significato analogo usate in registri più colloquiali: così, ad esempio, *effettuare* (*fare*), *suddividere* (*dividere*), *variare* (*cambiare*).

Alcune considerazioni si impongono sulla presenza reciproca di nomi e verbi del lessico accademico; se confrontiamo i dati dell'AIWL con quelli sulla percentuale media di nomi e verbi in corpora italiani scritti generali (Voghera, 2004), notiamo che, mentre la percentuale dei nomi accademici (47,3%) è più o meno equivalente, quella dei verbi (7,9) è pari a meno della metà degli altri corpora di italiano scritto.

Sembra dunque emergere che, a parte un numero molto ristretto di verbi, nomi e aggettivi siano le tipologie di vocaboli che caratterizzano maggiormente il lessico accademico. La forte densità lessicale della lingua scritta è del resto un dato consolidato (Halliday, 1985); nello specifico dello scritto accademico, la forte presenza di nomi è uno dei tratti linguistici che la caratterizzano molto nettamente come un genere testuale informativo, non narrativo, esplicito e astratto (Biber, 1988).

4.2. Collocazioni: tipologie e confronto con altri corpora

La collocazioni inserite nell'AIWL sono caratterizzate da diversi livelli di coesione interna, spaziando all'interno di un *continuum* che va da espressioni fisse e non interrompibili, come *arco di tempo*, a combinazioni di parole che tendono a co-occorrere in modo convenzionale, e con un grado minore di fissità, come *attribuire un significato*.

Confrontando le collocazioni accademiche con quelle di corpora di generi testuali diversi emergono chiaramente delle specificità relative alla comunicazione accademica. Il confronto sistematico con una lista di collocazioni verbo-articolo-nome di genere giornalistico e letterario, ad esempio (Spina, 2010), evidenzia alcune combinazioni che caratterizzano esclusivamente il corpus accademico (*porre le basi, introdurre il concetto, attribuire un significato*, *costituire la base*, *dare una definizione*, *porre le premesse*, *affrontare un tema*, *applicare un metodo*).

A una prima analisi, si riscontra anche una tendenza alla specializzazione di alcune tipologie di collocazioni; dai dati finora analizzati emerge che le combinazioni più "tipicamente accademiche" sono proprio quelle formate da verbo-articolo-nome, mentre la sequenza nome-preposizione-nome, che è anche quella più cristallizzata e meno soggetta a variazione interna, dà luogo frequentemente a combinazioni di parole assimilabili a tecnicismi e dunque, data la bassa dispersione nel corpus, molto spesso escluse dalla lista di frequenza. Citiamo solo alcuni esempi: *linguaggio di programmazione*, *controllo di gestione*, *variazione di temperatura*.

Uno studio più approfondito andrebbe sicuramente dedicato alla connessione tra le collocazioni accademiche con particolari funzioni testuali-informative tipiche dello scritto accademico (Paquot, 2007).

5. Conclusioni

In questo articolo è stata descritta la Lista di frequenza dell'italiano accademico (AIWL). La lista, estratta da un corpus bilanciato di italiano scritto accademico di 1 milione di parole (*Academic Italian Corpus*, o AIC), comprende 611 entrate lessicali, di cui 403 sono parole singole e le rimanenti 208 sono combinazioni di più di una parola. Gli obiettivi principali che sono alla base della costituzione dell'AIWL sono due:

- contribuire ad una descrizione più approfondita del lessico accademico, quall'insieme di unità lessicali non tecniche che caratterizzano la comunicazione accademica, a prescindere dall'area disciplinare;
- costituire una risorsa lessicografica e computazionale che funga da supporto al rafforzamento del lessico accademico di studenti universitari di lingua madre diversa dall'italiano e che sia integrabile in un ambiente di apprendimento di rete.

Rispetto ad analoghe liste di parole esistenti (Coxhead, 2000), l'AIWL ha due aspetti caratteristici, ed accoglie in questo senso le indicazioni contenute in Paquot (2007):

- include entrate lessicali singole e combinazioni di parole;
- è caratterizzata in senso produttivo oltre che ricettivo.

La lista di frequenza, ottenuta come risultato finale di operazioni di selezione basate su criteri statistici, è stata integrata da informazioni aggiuntive inserite in un *database* lessicale; tali informazioni, di tipo statistico-quantitativo, lessico-semantico e sintattico, sono mirate al trattamento automatico del lessico accademico. In particolare, il *database* lessicale sarà integrato in un ambiente di rete per l'apprendimento linguistico e fungerà da supporto per alcune applicazioni didattiche (selezione automatica dei lemmi accademici a partire da testi scritti, strumenti di scrittura assistita, generatori automatici di test *cloze*, per fare alcuni esempi), mirate al potenziamento della competenza del lessico accademico da parte di studenti non italofoni.

Riferimenti

Bauer L. and Nation I.S.P. (1993). Word families. *International Journal of Lexicography,* vol. 6, 4: 253-279.

Biber D. (1988). *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber D. (2004). Lexical bundles in academic speech and writing. In Lewandowska-Tomaszczyk, B., editor, *Practical Applications in Language and Computers (Proceedings of PALC 2003)*, Peter Lang: 165-178.

Bortolini U., Tagliavini C. and Zampolli A. (1971). *Lessico di frequenza della lingua italiana contemporanea*. Milano: Garzanti.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Cowie A.P. (1997). Phraseology in formal academic prose. In Aarts, J., de Mönnink, I. and Wekker, H., editors, *Studies in English Language and Teaching In Honour of Flor Aarts*, Amsterdam: Rodopi, pp. 43-56.

Coxhead A. (2000). A new academic word list. *TESOL Quarterly*, 34: 213-238.

Christ O. (1994). A modular and flexible architecture for an integrated corpus query system. In *COMPLEX '94*, Budapest.

De Mauro T. (1980). *Guida all'uso delle parole*. Torino: Editori Riuniti.

De Mauro T. (1999). *Gradit. Grande Dizionario Italiano dell'Uso*. Torino: Utet.

Granger S. and Paquot M. (2008). Disentangling the phraseological web. In Granger, S. and Meunier, F., editors, *Phraseology. An interdisciplinary perspective*. Amsterdam: John Benjamins, pp. 27-49.

Halliday M.K. (1985). *Spoken and written language*. Oxford: Oxford University Press.

Iengo S. (2009). *La competenza del lessico accademico italiano degli studenti non italofoni dell'Università per Stranieri di Perugia*. Tesi di laurea specialistica, Università per Stranieri di Perugia, http://hdl.handle.net/2447/99.

Jie Z. (2009). *Le difficoltà degli studenti universitari cinesi. Riflessioni dal punto di vista culturale e linguistico*. Tesi di laurea specialistica, Università per Stranieri di Perugia.

Nation I.S.P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nesselhauf N. (2005). *Collocations in a learner corpus*. Amsterdam: John Benjamins.

Oakes M. (1998). *Statistics for corpus linguistics*. Edinburgh University Press.

Oakey D. (2002). Formulaic language in English academic writing: A corpus-based study of the formal and functional variation of a lexical phrase in different academic disciplines. In Reppen, R., Fitzmaurice, S.M. and Biber, D., editors, *Using Corpora to Explore Linguistic Variation*, London: Longman, pp. 111-129.

Paquot M. (2007). Towards a productively-oriented academic word list. In Walinski, J., Kredens, K. and Gozdz-Roszkowski, S., editors, *Corpora and ICT in Language Studies. Proceedings of PALC 2005*. (Lodz Studies in Language 13), Peter Lang, pp. 127-140.

Schmid H. (1994). Probabilistic part-of-speech tagging us- ing decision trees. In *Proceedings of the International Conference on New Methods in Language Processing* (http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf).

Spina S. (forthcoming,a). Building a suite of online resources to support academic vocabulary learning. In EUROCALL 2009 conference, Universidad Politécnica de Valencia (Valencia, September 2009).

Spina S. (forthcoming,b). *La competenza del lessico accademico da parte di studenti non italofoni*. Manuscript submitted for publication.

Spina S. (2010). The Dici Project: towards a Dictionary of Italian Collocations integrated with an online language learning platform. In *Proceeding of eLexicography in the 21st century: new challenges, new applications*, Louvain-La-Neuve, 22-24 octobre 2009, Cahiers du Cental.

Voghera M. (2004). La distribuzione delle parti del discorso nel parlato e nello scritto. In Van Deyck, R., Sornicola, R. and Kabatèk, J., editors, *La variabilité en langue, I. Langue parlée et langue écrite dans le présent et dans le passé, II. Les quatre variations, Communication & Cognition*. Ghent University, pp. 261-284.

Vongpumivitch V., Huang J. and Chang Y. (2009). Frequency Analysis of the Words in the Academic Word List (AWL) and Non-AWL Content Words in Applied Linguistics Research Papers. *English for Specific Purposes*, vol. 28, n. 1: 33-41.

West M. (1953). *A General Service List of English Words*. London: Longman.