

LE COLLOCAZIONI N-ADJ DELL'ITALIANO ACCADEMICO PARLATO. UN'INDAGINE CORPUS-BASED

Diana Peppoloni¹

1. INTRODUZIONE²

I recenti progressi nella linguistica dei corpora e lo sviluppo di grandi corpora elettronici specializzati consentono agli esperti di esplorare nuove tecniche (linguistiche, statistiche, ecc.) per gestire la loro fonte primaria di informazione. Molti studi basati sui corpora mirano a identificare modelli lessicali e grammaticali ricorrenti, legati al fenomeno linguistico delle *collocazioni*.

Questo contributo descrive la creazione di una risorsa originale, un corpus di piccole dimensioni della lingua italiana accademica parlata, utilizzato come base di dati per l'estrazione automatica delle collocazioni in questo tipo di contesti comunicativi, e l'ulteriore processo di analisi e validazione di queste combinazioni lessicali. L'ipotesi di ricerca si basa su precedenti lavori incentrati su diversi fenomeni fraseologici della lingua inglese (Biber, 2006; Coxhead, 2000), ma allo stesso tempo si differenzia da essi, per la sua attenzione specifica alle collocazioni e per la lingua di destinazione, ovvero l'italiano. L'attenzione ricade su una specifica struttura collocativa, costituita dai componenti *nome* e *aggettivo* (N-Adj), collegati in modo preferibile e ricorrente rispetto ad altre possibili combinazioni, per veicolare un significato peculiare complessivo.

La categoria di collocazioni di nostro interesse è stata estratta dal corpus di riferimento ASIC (*Academic Spoken Italian Corpus*) utilizzando le funzionalità di Sketch Engine (Kilgarriff *et al.*, 2004), che utilizza le informazioni derivanti dalle precedenti operazioni di POS-tagging effettuate in fase di compilazione del corpus, per rintracciare tutte quelle coppie di parole costituite dai *tag* grammaticali indicati in fase di ricerca.

L'esito di questo procedimento di calcolo è stato un elenco di oltre 600 combinazioni di parole N-Adj, con una frequenza di occorrenza pari o superiore a 3. Su questo campione di dati iniziale è stata fatta una prima procedura di scrematura, escludendo quelle che erano combinazioni casuali della varietà linguistica considerata. Si è arrivati così a un campione ristretto di 273 possibili collocazioni. Trattandosi di comunicazione accademica, è stato necessario applicare un ulteriore metodo di filtraggio dei dati; difatti, perché le collocazioni individuate rappresentassero quel piano intermedio della comunicazione trasversale a diverse discipline e non fossero idiosincratiche di una singola area del sapere, si è reso necessario selezionare solo le combinazioni distribuite in almeno due delle aree disciplinari comprese in ASIC. Da questa operazione è risultato un campione di 56 possibili collocazioni della lingua italiana accademica parlata.

¹ Università degli Studi di Perugia.

² Lo studio è promosso dal Centro di Ricerca Interuniversitario “LinE - Language in Education” (www.languageineducation.eu) nell'ambito del cluster scientifico “La comunicazione accademica parlata: analisi e strumenti dalla prospettiva della glottodidattica”.

This study is promoted by the Inter-university Research Centre “LinE – Language in Education” (www.languageineducation.eu) within the scientific cluster “The Spoken Academic Communication: Analysis and Resources from the Educational Linguistics’ Perspective”.

Per validare tali dati, in una fase successiva sono stati coinvolti 11 esperti linguistici, docenti universitari di madrelingua italiana, che avessero un'elevata competenza linguistica e una familiarità con la comunicazione accademica. Attraverso la compilazione di un formulario appositamente predisposto, ognuno di loro ha valutato un gruppo di cinque o sei combinazioni lessicali, cui andava attribuito un giudizio positivo o negativo di collocabilità; in caso affermativo, veniva inoltre richiesto di definire il significato delle combinazioni di parole assegnate, utilizzando fino a 140 caratteri e fornendo, eventualmente, degli esempi di possibili contesti d'uso.

Gli studi corpus-based si prestano a rispondere a molteplici tipi di quesiti di ricerca (Biber *et al.*, 1998); in questo studio ne abbiamo considerati principalmente due, ovvero la verifica dei contesti d'uso e delle caratteristiche di una particolare tipologia di espressioni polirematiche, ovvero le collocazioni, e la correlazione che queste intrattengono con un determinato registro linguistico, ovvero quello dell'oralità. Dunque il tema del contributo è duplice, da un lato riguarda l'analisi di fenomeni combinatori del lessico, dall'altro cerca di descrivere aspetti legati a una specifica varietà linguistica quale quella che caratterizza la comunicazione accademica parlata.

Il focus sul fenomeno delle collocazioni si deve all'importanza che queste rivestono nel percorso acquisitivo degli apprendenti di lingua, poiché la conoscenza e il corretto impiego di tali unità fraseologiche sono indice di un livello avanzato di competenza linguistico-comunicativa del parlante e della sua eventuale appartenenza a un certo gruppo sociale o professionale, come nel caso della comunicazione accademica (Wray, 2002). Inoltre, un uso estensivo delle collocazioni è una delle evidenze più significative nel distinguere le produzioni dei parlanti nativi da quelle dei non-nativi, che tendono a sottoutilizzare o utilizzare in maniera impropria tali *chunks* linguistici (Wray, Perkins, 2000).

Le collocazioni costituiscono un ostacolo anche per gli apprendenti di livello avanzato, per varie ragioni. Secondo quanto afferma Nation (2001) ciò dipende dall'imprevedibilità della combinatoria degli elementi che le compongono. Laufer (2011) conferma questa posizione, specificando come l'uso delle collocazioni risulti problematico per gli apprendenti stranieri, indipendentemente dagli anni di istruzione ricevuta nella lingua target, dalla loro lingua materna o dal tipo di attività che devono svolgere. In particolare, è l'uso produttivo delle collocazioni a rivelarsi particolarmente problematico per gli apprendenti stranieri.

Sebbene l'acquisizione delle collocazioni sia una componente imprescindibile dello sviluppo della competenza linguistico-comunicativa dell'apprendente, essa viene spesso sottostimata nella classe di lingua. Ciò si deve spesso a una carenza di materiali e risorse specificatamente dedicati, da cui la necessità di questo studio.

2. LO SPAZIO DELLA COMUNICAZIONE ACCADEMICA

Su un piano intermedio, sito tra il vocabolario tecnico e quello generale di una lingua, si trovano quelle unità lessicali ed espressioni combinatorie che vengono utilizzate in diversi spazi della comunicazione accademica, in maniera indipendente dai contenuti disciplinari veicolati (Farrell, 1990). Tali elementi della lingua costituiscono il vocabolario della lingua accademica (Nation, 2001).

Per lingua accademica si intende quella «specialized language, both oral and written, of academic environments that facilitates communication and reflection on disciplinary contents» (Nagy, Townsend, 2012: 92). Questa varietà linguistica viene prodotta per scopi comunicativi e un pubblico specifici, e utilizzata in un contesto circoscritto, ovvero quello universitario, per acquisire nuove conoscenze o comprendere nuovi contenuti, per

formulare ipotesi e argomentazioni, nonché per comunicare queste nozioni esternamente, a un pubblico di carattere generalista (Bailey, Heritage, 2008; Schleppegrell, 2004; Facione, 1990).

Già Vygotsky (1987) evidenzia una distinzione cruciale tra il linguaggio acquisito da un bambino durante la prima infanzia e quello che un individuo deve sviluppare per orientarsi con successo nel mondo accademico e, successivamente, nel contesto professionale. L'utilizzo della lingua in ambiti accademici e lavorativi richiede l'aderenza a un insieme di regole complesse, esplicite e spesso difficili da comprendere per chi ne è esterno. Queste norme non riguardano soltanto gli aspetti linguistici, ma sono strettamente legate alle dinamiche proprie delle istituzioni accademiche e dei diversi settori disciplinari. Chiunque voglia integrarsi efficacemente in tali contesti necessita di padroneggiare la lingua accademica, strumento imprescindibile di comunicazione e acquisizione di nuove conoscenze.

Se si considera l'assunto di Bourdieu e Passeron (1994: 8) secondo cui «academic language is no one's mother tongue», appare chiaro come sia per gli studenti e gli esperti parlanti nativi della lingua in uso nella comunità accademica di appartenenza, che per quelli stranieri, tale varietà si configuri come un sistema con una tendenza evolutiva, dal punto di vista della complessità e della competenza. Difatti, la capacità di gestire accuratamente e appropriatamente la lingua accademica aumenta man mano che si progredisce nel proprio percorso universitario, parallelamente allo sviluppo della conoscenza dei contenuti disciplinari e degli specifici marcatori linguistici utili a esprimerli e propri delle diverse aree di contenuto (Hyland, 2002; Anstrom *et al.*, 2010; Hamp-Lyons, 2011). Coerentemente con questa visione, Jordan (1997: 5) descrive il linguaggio accademico come «a general academic [...] register, incorporating a formal, academic style, with proficiency in the language use».

La padronanza della lingua accademica nazionale incoraggia la comunità scientifica a resistere all'influenza pervasiva dell'inglese e di altre lingue internazionali. Questo impatto non si limita solo ad aspetti di natura linguistica, poiché, come sostiene Marazzini (2021), l'italiano sembra oggi meno adatto ad argomentare su temi scientifici, ma tocca anche i paradigmi scientifici che vengono costruiti attraverso la lingua stessa. Non si tratta solo di terminologia, ma soprattutto di come le etichette terminologiche vengono combinate per costruire modelli, teorie e metodologie, modellandosi sulla lingua che li descrive, avvantaggiando i parlanti nativi di quella lingua. L'abbandono dei sistemi linguistici nazionali in favore di una lingua franca dell'accademia rischia di relegarli a una posizione periferica nel panorama scientifico globale. Se il discorso accademico deve essere formulato in un'unica lingua sovranazionale per essere considerato rilevante e fruibile, sarà quella la lingua su cui si articoleranno i quadri epistemologici e terminologici del sapere, mentre le altre lingue riceveranno tali concetti solo di riflesso, senza aver sviluppato le parole, o meglio i termini, per poterli diffondere e discutere.

Perché gli esperti possano utilizzare la loro lingua materna nei loro lavori, ovvero il sistema linguistico dove godono del potenziale espressivo più efficace, devono verificarsi almeno due condizioni, una riguardante la lingua stessa e l'altra il parlante che la usa. Innanzitutto, la lingua utilizzata deve godere di prestigio e rilevanza internazionale e ciò è possibile solo se i maggiori esponenti del panorama accademico di riferimento scelgono di utilizzarla per diffondere i propri studi, sia a un pubblico specialistico che generalista. In secondo luogo, tali esperti devono essere formati su quali siano le caratteristiche strutturali e retoriche di quella varietà linguistica.

Questo studio intende descrivere il fenomeno delle collocazioni N-Adj nella varietà linguistica accademica parlata della lingua italiana, andando a verificarne l'incidenza e l'uso nei contesti comunicativi e disciplinari contenuti nel corpus ASIC. La conoscenza di questi fenomeni promuove la formazione di interlocutori consapevoli e capaci di

«understand words that are used to talk and write about science» (Snow, 2008: 452), vale a dire di padroneggiare un vocabolario sofisticato e astratto, che includa parole accademiche multiuso insieme a quelle specifiche della disciplina oggetto di studio e di ricerca.

3. VERSO UNA DEFINIZIONE DI COLLOCAZIONE

Le espressioni polirematiche, di cui le collocazioni costituiscono un sottogruppo, vengono definite nell'Enciclopedia Treccani³ come combinazioni di parole «formate da più elementi che costituiscono un insieme non scomponibile, il cui significato complessivo è autonomo rispetto ai singoli costituenti».

Si tratta di combinazioni lessicali sintatticamente articolate, che esprimono però un'unica unità semantica, non scomponibile nei significati dei singoli componenti che le costituiscono. Che si tratti di un significato unitario e non compositivo è facilmente percepibile quando tali composti lessicali hanno un'accezione figurata. Ad esempio, in italiano, possiamo utilizzare la collocazione *testa calda* per riferirci a un individuo impulsivo, dalle idee audaci, ma non è altresì accettabile utilizzare l'espressione *testa bollente* per enfatizzare o esasperare tale tratto connotativo della persona o, al contrario, per indicare un soggetto dall'animo pacifico, ricorrendo all'espressione *testa fredda*. La complessità sintattica di tali espressioni polirematiche può essere dimostrata, invece, provando a sostituire uno degli elementi che le compongono; se sostituissimo l'elemento *rischio* all'interno dell'espressione *rischio calcolato* con *pericolo*, ottenendo dunque *pericolo calcolato*, l'unità fraseologica considerata perderebbe la sua riconoscibilità, dettata non tanto da fattori interni al sistema linguistico, quanto a questioni di natura culturale, di uso e di abitudine (Tiberii, 2012: 3).

La nozione di collocazione è tradizionalmente associata al lavoro di Firth (1957: 106-107), promotore dello «study of key-words, pivotal words, leading words, by presenting them in the company they usually keep». Studiosi posteriori, ispirati dalla sua idea, hanno tentato di rendere la nozione meno oscura e più operativa, e di proseguire il suo studio utilizzando corpora testuali, dove veniva ben rappresentata l'idea del «meaning by collocation» (*ibidem*: 194). Fortemente influenzato dal lavoro di Malinowski (1921) sui fattori sociali e culturali inerenti all'uso e all'evoluzione della lingua, Firth considera il significato delle parole come una funzione del contesto in cui queste appaiono e si combinano tra loro in modo abituale, ma arbitrario, all'interno di ciascun sistema linguistico. Egli arriverà infatti ad affermare che «The basic assumption of the theory of analysis by levels is that any text can be regarded as a constituent of a context of situation» (Firth, 1957: 175).

Le collocazioni non sono semplicemente elementi fissi, ma il prodotto del processo dinamico di produzione linguistica da parte dei parlanti nativi di una certa lingua, che gradualmente si istituzionalizza attraverso l'uso in un dato ambiente linguistico. Ciò implica che diversi contesti situazionali, ovvero varietà nazionali o regionali, lingue speciali o generi specifici, svilupperebbero collocazioni tipiche di quel contesto comunicativo.

Jones e Sinclair (1974: 19) descrivono i fenomeni collocazionali come «regular collocation between items, such that they occur more often than their respective frequencies and the length of the text in which they occur would predict». Nel 1991 Sinclair elabora l'ormai ben noto *idiom principle*, per cui i parlanti nella loro L1 tenderebbero a utilizzare blocchi riutilizzabili di lingua all'interno di contesti specifici, i cosiddetti *chunks* linguistici. Già Pawley e Syder (1983: 205), prima di lui, avevano affermato che «a minority

³ [https://www.treccani.it/enciclopedia/espressioni-polirematiche_\(La-grammatica-italiana\)/](https://www.treccani.it/enciclopedia/espressioni-polirematiche_(La-grammatica-italiana)/).

of spoken clauses are entirely novel creations in the sense that the combination of lexical items is new to the speaker», evidenziando la tendenza dei parlanti a riutilizzare combinazioni lessicali all'interno dei propri enunciati, che non si presentano dunque come produzioni creative totalmente inedite. Della stessa idea sarà successivamente Cowie (1988) che si riferirà alle collocazioni come a *ready-made expressions* o *prefabricated routines*. In questa prospettiva, l'uso della lingua è di natura fondamentalmente idiomatica.

Sinclair definisce le collocazioni su base puramente statistica, adottando una prospettiva di tipo *frequency-based*, come segnalato da Nesselhauf (2005). Egli (1991: 170) descriverà le unità collocazionali come «the occurrence of two or more words within a short space of each other in a text» con una frequenza inattesa. Naturalmente la frequenza nella combinazione di due unità lessicali non può essere l'unico parametro considerato, poiché non dà informazioni a sufficienza sulla struttura delle espressioni polirematiche derivanti.

Spostandoci di prospettiva, verso un altro tipo di approccio definitorio al concetto di collocazione (Nesselhauf, 2005), ovvero quello fraseologico, introduciamo un nuovo parametro, che è quello della restrizione lessicale. Sviluppando questo concetto, Jezek e Sprugnoli (2023: 40) definiscono una collocazione come una «combinazione di parole soggetta a restrizione lessicale, per cui la scelta di una specifica parola (il collocato) per esprimere un determinato significato, è condizionata da una seconda parola (che qui chiameremo base) alla quale questo significato è riferito». È il caso, ad esempio, di combinazioni come *vittoria schiacciante*, in cui la base, *vittoria*, si combina con il collocato, *schiacciante*, per veicolare il tratto semantico dell'intensità.

Il comportamento della base all'interno di una collocazione, ovvero la sua capacità di selezionare un collocato specifico tra una lista di possibili alternative, per esprimere un significato che non avrebbe se fosse combinata con un'unità diversa (e.g. *vittoria evidente* invece di *vittoria schiacciante*) richiama propriamente il già citato concetto firthiano, elaborato molto tempo prima, del «meaning by collocation».

Tiberii (2012), autrice del *Dizionario delle collocazioni. Le combinazioni delle parole in italiano*, nel definire le collocazioni sposta ulteriormente l'obiettivo, concentrandosi sull'arbitrarietà e la convenzionalità dei fenomeni collocazionali, che in virtù di queste caratteristiche variano sul piano interlinguistico, poiché determinati da fattori culturali e dall'abitudine all'uso che ne hanno i parlanti nativi di una data lingua. Ciò, come vedremo nella prossima sezione, rende le collocazioni uno degli ostacoli più significativi per gli apprendenti di lingua straniera, anche di livello avanzato, e, per questo, uno dei fenomeni che andrebbero affrontati con maggiore cura nei percorsi acquisizionali, sia in termini metodologici che di disponibilità di materiali e risorse adeguati a introdurli e trattarli.

3.1. *L'importanza delle collocazioni nell'acquisizione di una lingua straniera*

La competenza collocazionale è una componente chiave dello sviluppo della competenza linguistico-comunicativa degli studenti di livello avanzato di una lingua straniera (Wray, 2002); essa incarna una sfida didattica importante, riferendosi a una parte altamente stereotipata e convenzionale di una lingua. Nonostante la loro centralità nel percorso acquisizionale, le collocazioni vengono trattate solo marginalmente nella classe di lingua, soprattutto a causa della mancanza di risorse specifiche dedicate. Per questo motivo il presente studio intende indagare tale fenomeno nello spazio della lingua accademica parlata, ed elaborare materiali specifici, che possano supportare docenti e apprendenti nel percorso di insegnamento e apprendimento dell'italiano accademico. Seguendo l'idea di Bourdier e Passeron (1994: 8) per cui «academic language is [...] no one's mother tongue», si intuisce come tale varietà linguistica si configuri come un sistema

con una tendenza evolutiva, dal punto di vista della complessità e della competenza, sia per gli studenti e gli esperti parlanti nativi della lingua in uso nella comunità accademica di appartenenza, che per quelli stranieri, e debba dunque essere studiata esplicitamente da entrambe queste categorie di soggetti.

Riprendendo le parole di Tiberii (2012: 3), «Una elevata competenza linguistica è caratterizzata [...] non solo da conoscenze grammaticali e sintattiche, ma soprattutto dalla capacità di utilizzare le combinazioni lessicali *proprie* della lingua», diviene immediatamente comprensibile come le collocazioni giochino un ruolo fondamentale nel determinare il livello di competenza linguistico-comunicativa in una data lingua, nonché l'appartenenza di un parlante a una determinata comunità linguistica o a un gruppo sociale o professionale, come nel caso della comunicazione accademica. Come sostiene Grace (1987), parlare una lingua in modo idiomatico significa conformarsi a modi di esprimersi prestabiliti e largamente condivisi dai parlanti nativi. Difatti, nonostante una certa variabilità intrinseca nella loro formulazione (Fernando, 1996; Corpas Pastor, 2003; Masini, 2009), i fraseologismi vengono percepiti come modalità espressive convenzionali all'interno di una determinata comunità linguistica e quindi culturale. Sapere come manipolarli e utilizzarli appropriatamente rientra nella competenza di un parlante nativo (Fillmore *et al.*, 1988; Nattinger, DeCarrico, 1992).

Cowie (1992: 10) afferma che «the sheer density of ready-made units in various types of written text is a fact that any approach to the teaching of writing to foreign students has to come with». Se questo è vero per il registro scritto, ha una valenza ancor più evidente nel parlato, dove il ritmo dell'eloquio è più elevato e incontrollabile e l'efficacia pragmatica delle produzioni linguistiche ancor più necessaria e immediata.

Nonostante l'acquisizione di formule conversazionali durante il percorso di acquisizione di una lingua straniera costituisca un fattore centrale nello sviluppo della competenza linguistica di un apprendente, influenzando aspetti come la fluenza e l'efficacia pragmatica dei suoi enunciati, la maggior parte degli apprendenti di lingua straniera incontra notevoli difficoltà nel decodificare e produrre espressioni collocazionali, per loro natura altamente convenzionali e idiosincratiche del sistema linguistico in cui vengono prodotte. Ciò è vero non solo per gli apprendenti principianti, ma anche per quelli di livello intermedio e avanzato, costituendo un elemento distintivo del parlante non nativo di una data lingua (Wray, 2002; Nattinger, DeCarrico, 1992). Quest'ultimo, infatti, fa un uso molto più circoscritto di tali unità fraseologiche rispetto ai parlanti nativi, commettendo in ogni caso significativi errori strutturali e contestuali e alternando espressioni appropriate, inappropriate e interlinguistiche, queste ultime frutto di inevitabili processi di transfer linguistico dalla propria lingua materna (Wray, Perkins, 2000).

Le collocazioni, dunque, in quanto unità linguistiche, semantiche e culturali rappresentano una componente imprescindibile per lo sviluppo di competenze comunicative avanzate per gli apprendenti di lingua straniera.

In prospettiva glottodidattica, però, gli idiomatismi spesso non vengono introdotti con la dovuta attenzione dal docente di lingua, soprattutto a causa dell'assenza di efficaci indicazioni metodologiche e della ridotta disponibilità di risorse dedicate.

Sulla base di queste riflessioni, questo contributo propone un'analisi di carattere linguistico, che ambisce a fornire strumenti glottodidattici, colmando alcune attuali lacune strumentali riguardanti specifici fenomeni fraseologici, quali le collocazioni, che incidono sullo sviluppo strutturale e comunicativo dell'apprendente (Yorio, 1989). A tale scopo è stata dunque elaborata una lista di collocazioni N-Adj, automaticamente estratta dal corpus ASIC e validata da parlanti nativi esperti da un punto di vista accademico e linguistico. Il docente potrà utilizzare tale strumento per elaborare testi, prove valutative, attività di esplorazione diretta e indiretta del corpus di riferimento e dei contesti d'uso delle espressioni selezionate. Egli disporrà, insomma, di materiale originale,

contestualizzato, pronto per essere utilizzato nella classe di lingua, a integrazione dell'indirizzo metodologico applicato nel corso di lingua.

4. MATERIALI E METODI

4.1. *Il corpus ASIC*

Per indagare il fenomeno delle collocazioni N-Adj nello spazio della comunicazione accademica italiana parlata, è stato condotto uno studio corpus-based, analizzando un'ampia gamma di produzioni autentiche provenienti dal registro orale, contenute all'interno del corpus ASIC (Academic Spoken Italian Corpus)⁴.

Il corpus ASIC è stato costruito bilanciando i dati in due direzioni; orizzontalmente, suddividendo le aree tematiche di appartenenza dei testi raccolti (umanistica/ giuridico-economica/scientifica/medica), e verticalmente, diversificando le situazioni comunicative considerate rilevanti per il progetto (lezioni/esami orali/convegni/workshop).

Pertanto i dati raccolti sono riferiti a diverse tipologie di conoscenze scientifiche e a diversi contesti interazionali, tipici della comunicazione accademica; tale operazione di bilanciamento è stata necessaria per soddisfare un parametro fondamentale nella costruzione di un corpus, ovvero il bilanciamento dei dati in esso contenuti (Mc Enery, Hardie, 2011). Dato che un corpus dovrebbe costituire un campione rappresentativo di una porzione della lingua considerata, l'accuratezza nella selezione dei testi di cui si compone influisce in maniera significativa sulla sua validità. Con l'aggettivo *rappresentativo* si fa riferimento alla capacità di un corpus di tenere traccia dell'intero complesso delle caratteristiche strutturali e delle proprietà combinatorie e contestuali di una data varietà linguistica, consentendo ai ricercatori di estendere tali caratteristiche linguistiche all'intera popolazione linguistica. Quindi la struttura dei dati è fondamentale per costruire un supporto concreto alle teorie e alle analisi degli esperti; in altre parole, la validità di una ricerca dipende dalla validità dei dati raccolti e anche dalla loro corretta e logica organizzazione.

La prima fase di realizzazione del Corpus ASIC è stata, dunque, la raccolta dei dati, effettuata per mezzo della registrazione audio di lezioni, esami, conferenze e workshop di vario argomento, nel periodo compreso tra dicembre 2022 e giugno 2024. Fatta eccezione per alcune giornate di studio tenutesi in altri Atenei italiani, a cui è stato possibile partecipare a distanza, i dati sono stati interamente rilevati in presenza, presso l'Università degli Studi di Perugia, nei Dipartimenti di Lettere, Giurisprudenza, Economia e Commercio, Ingegneria, Veterinaria e Medicina e Chirurgia.

La scelta dei Dipartimenti selezionati per il progetto di ricerca è stata guidata, come già sottolineato, dall'esigenza di poter disporre di un corpus capace di rappresentare la varietà linguistica tipica dell'ambiente universitario, indipendentemente dalla specificità dei diversi settori scientifici considerati. Quattro sono le macro-aree disciplinari individuate per la registrazione dei dati: 1) area umanistica 2) area giuridico-economica 3) area scientifica 4) area medica. Una volta registrati i vari eventi comunicativi (i.e. lezioni, esami, convegni, workshop e seminari), i dati sono stati trascritti e il corpus consta attualmente di 333.402 entrate. Di seguito, riportiamo il prospetto sintetico ed esplicativo dello stato dell'arte del progetto, con l'indicazione del numero di entrate per ciascuna area disciplinare di riferimento e per le diverse situazioni comunicative (Tabella 1).

⁴ Il corpus ASIC è stato sviluppato all'interno del progetto di ricerca DACA (Descrizione e Analisi della Comunicazione Accademica), finanziato dai Fondi di Ricerca di Ateneo 2023.

Tabella 1. *Insieme dei dati contenuti nel corpus ASIC ripartiti per ambiti disciplinari e situazioni comunicative*

Area disciplinare	Situazione comunicativa	Entrate
Umanistica	Lezioni	52274
	Convegni e seminari	47949
	Esami	10.435
Totale		110.658
Area disciplinare	Situazione comunicativa	Entrate
Giuridico-economica	Lezioni	54373
	Convegni e seminari	21.488
	Esami	(in fase di trascrizione)
Totale		75.859
Area disciplinare	Situazione comunicativa	Entrate
Scientifica	Lezioni	62.126
	Convegni e seminari	(in fase di trascrizione)
	Esami	(in fase di trascrizione)
Totale		62.126
Area disciplinare	Situazione comunicativa	Entrate
Medica	Lezioni	64.221
	Convegni e seminari	(in fase di trascrizione)
	Esami	20.538
Totale		84.759
Totale complessivo delle entrate:		333.402

4.2. *Le collocazioni N-Adj della lingua accademica parlata*

I metodi più comuni per l'estrazione automatica di combinazioni collocazionali prevedono che si effettui una ricerca in un corpus bilanciato e rappresentativo della varietà linguistica analizzata, tramite l'uso di *tagset* grammaticali e, successivamente, un'ulteriore scrematura degli abbinamenti rilevati in base alla loro frequenza o a varie misure di associazione, al fine di distinguere le combinazioni significative dalle sequenze di parole che non costituiscono alcun tipo di unità fraseologica rilevante.

In generale, la suddetta ricerca può essere eseguita utilizzando due diverse tipologie di legami tra le parole, ovvero i *pattern* morfosintattici superficiali (POS-tag) (approccio *P-based*) o relazioni di dipendenza sintattica (approccio *S-based*). I primi richiedono che si disponga di un corpus con *POS-tag*gato, in cui a ciascuna delle entrate del corpus venga attribuita una parte del discorso (i.e. Noun, Verb, Adjective, etc.), a partire da un elenco di possibili *Part Of Speech-tag* considerati rappresentativi in una determinata lingua. I secondi, d'altro canto, richiedono un corpus su cui siano state eseguite procedure di *parsing*, ossia di identificazione di relazioni sintattiche significative.

In questo studio, adottiamo un metodo *P-based*, possibile grazie alle operazioni di *POS-tagging* preliminarmente effettuate su ASIC tramite il software Sketch Engine. Tale software consente agli utenti, tra le numerose funzioni disponibili, di esaminare e

manipolare automaticamente i dati in un corpus, eseguendo operazioni come l'estrazione automatica di collocazioni e la creazione di elenchi di frequenza di parole.

L'estrazione delle collocazioni corrispondenti alla sequenza POS N-Adj è stata eseguita utilizzando la seguente query: [tag = "N. *"] [tag = "ADV.*"]? [tag = "A. *"].

Secondo questa espressione regolare, il primo elemento della ricerca automatica è una parola contrassegnata nel corpus come sostantivo, eventualmente seguita da un avverbio (e.g. *teoria davvero innovativa, modo molto diverso*), mentre l'ultimo elemento sarà una parola etichettata come aggettivo. Questa espressione regolare sviluppata nel linguaggio di programmazione CQL, ci consente l'estrazione automatica di diverse possibili combinazioni N-Adj della lingua italiana accademica parlata, che vedono anche il possibile fraporsi di un'ulteriore parte del discorso.

Sebbene un approccio *P-based*, come quello qui adottato, fornisca risultati accurati, esso presenta anche alcune possibili ambiguità. Ad esempio il *tagger* potrebbe assegnare alle entrate del corpus delle etichette grammaticali errate nel contesto frasale in cui queste compaiono; ad esempio, potrebbe identificare come aggettivo un'entrata che ha come omonimo un participio passato, come nel caso di *problema legato* [freq. 4].

Proprio per evitare di includere sequenze errate dal punto di vista del *tagging*, il ricercatore interviene ex-post per verificare che le combinazioni estratte siano effettivamente sequenze N-Adj, in questo caso. Quindi, sebbene si tratti di una procedura di estrazione automatica, complessivamente la selezione del materiale comporta un intervento soggettivo dell'esperto, trasformando dunque l'operazione in semi-automatica. Tuttavia, l'intervento umano è successivo all'estrazione dei dati, quindi non si verifica una manipolazione aprioristica degli stessi.

Dopo aver eseguito l'estrazione dei dati e la verifica della correttezza delle sequenze taggate, abbiamo ottenuto un primo elenco di oltre 640 collocazioni, che è stato poi ulteriormente filtrato.

Sono stati cioè anzitutto rimossi tutti quei candidati con frequenza = 0 < di 3, per ridurre il fattore statistico della casualità. Inoltre, sono state rimosse tutte quelle combinazioni esplicitamente specialistiche, cioè proprie di un settore disciplinare e non del più generale livello rappresentato dalla comunicazione accademica, come *storia filosofica* (freq. 46) un tecnicismo ad alta frequenza del settore umanistico, *azione amministrativa* (freq. 46) tecnicismo ad alta frequenza del settore giuridico-economico, *corpo calloso* (freq. 44) tecnicismo ad alta frequenza dell'area medica e *espressione regolare* (freq. 81) tecnicismo ad alta frequenza di ambito scientifico. Inoltre, sono state eliminate dalla lista delle possibili collocazioni estratte, anche quelle combinazioni che includevano un nome proprio di persona, come *Federico secondo*, o relative alla denominazione di un'istituzione, come *Unione Europea* o *Casa Bianca*, nonché i forestierismi, come *target like*.

Così facendo, la lista si è ridotta alle 273 collocazioni N-Adj più frequenti presenti nel corpus ASIC. Un campione dei dati ottenuti è descritto nella Tabella 2⁵.

Tabella 2. *Campione delle collocazioni N-Adj estratte insieme alla loro frequenza di occorrenza in ASIC*

Lemma	Frequenza
interesse pubblico	60
vertice politico	47
potere amministrativo	39
procedura remoto	38

⁵ La raccolta completa dei dati estratti e filtrati è disponibile nell'Appendice 1.

gruppo funzionale	35
organo politico	33
sistema operativo	32
mano sinistra	31
atto amministrativo	24
indirizzo politico	24
ente pubblico	23
forma vera	21
volta scorsa	20
potere discrezionale	19

Dovendo rappresentare le collocazioni N-Adj proprie della comunicazione accademica parlata, cioè di quel livello linguistico intermedio compreso tra la lingua comune e la lingua specialistica, propria dei diversi ambiti disciplinari presenti in ASIC, è stato necessario filtrare ulteriormente i dati. Tale procedura di filtraggio ha permesso di individuare le collocazioni effettivamente disperse nel corpus analizzato, ovvero quelle presenti in almeno due dei quattro settori disciplinari da cui provenivano i dati raccolti. In questo modo, le collocazioni rimanenti risultano effettivamente appartenere al piano trasversale della comunicazione accademica e rappresentare unità fraseologiche utilizzate nei diversi discorsi propri di questo contesto.

Il risultato di questo ulteriore filtro applicato è una lista di 56 collocazioni, di cui si propone un estratto nella Tabella 3⁶.

Tabella 3. Estratto della lista delle 56 collocazioni filtrate in base alla distribuzione nel corpus ASIC

Lemma	Frequenza	Area disciplinare	Area disciplinare	Area disciplinare	Area disciplinare
1. sistema operativo	32	Giurecon		Scient	
2. ente pubblico	23	Giurecon			Uman
3. volta scorsa	20		Med	Scient	Uman
4. tempo stesso	12	Giurecon	Med	Scient	Uman
5. essere umano	12	Giurecon	Med		Uman
6. modo diverso	12	Giurecon	Med		Uman
7. intelligenza artificiale	11	Giurecon			Uman
8. brano musicale	10	Giurecon	Med		

È interessante notare come alcune delle collocazioni rintracciate si inseriscano negli effettivi contesti d'uso in locuzioni più ampie, come ad esempio *costo zero* (freq. 3), che ricorre all'interno del corpus nell'espressione *a costo zero*. Nell'ottica dell'acquisizione di tali combinazioni, è importante che il parlante ne conosca non solo gli aspetti semantici che le caratterizzano, ma anche la struttura sintattica che è loro propria. La possibilità di risalire

⁶ L'elenco completo delle 56 collocazioni filtrate è disponibile nell'Appendice 2.

ai contesti d'uso in cui tali collocazioni compaiono, permette all'apprendente e al docente, di verificare e lavorare, direttamente o indirettamente, su questi aspetti della lingua.

La lista finale delle collocazioni estratte può essere intesa come un elenco di collocazioni empiriche, vale a dire quelle combinazioni di parole i cui componenti co-occorrono frequentemente (in questo caso la soglia è 3). Le collocazioni empiriche (Evert, 2008) sono quelle combinazioni di parole, contigue o meno, che si presentano insieme più spesso del previsto (Krishnamurthy, 2006). La tendenza di due o più parole a comparire congiuntamente può indicare che fanno parte di un'espressione rigida o meno libera, rispetto ad altre combinazioni lessicali che si adattano facilmente insieme a molte altre parole in modo casuale. In questo caso, quindi, potremmo considerare la frequenza di co-occorrenza come una misura in grado di identificare, nell'insieme delle innumerevoli co-occorrenze del corpus, quelle che più probabilmente costituiscono dei fenomeni fraseologici. È facile verificare, tuttavia, che un elenco di espressioni ordinate in ordine decrescente di co-occorrenza, pur includendo espressioni che appartengono all'insieme delle collocazioni, selezionerà anche espressioni generalmente considerate dai parlanti nativi come libere, cioè non ristrette semanticamente e/o sintatticamente. Questo è il caso, nel nostro studio, di coppie di parole come *cosa diversa* o *parte superiore*. Considerata la presenza di tali espressioni, si è deciso di affidarsi alla competenza metalinguistica di parlanti nativi italiani altamente competenti dal punto di vista linguistico e familiari rispetto allo spazio della comunicazione accademica, per valutare quali delle coppie di parole estratte automaticamente potessero essere considerate effettive collocazioni della lingua italiana accademica parlata, come illustrato nella sezione seguente.

4.3. Validazione del campione di dati filtrato

Poiché le collocazioni sono fenomeni lessicali complessi che, per loro stessa natura, sfuggono ai tentativi di formalizzarli in insiemi di regole ben definiti, rigidi e sempre validi, non tutte le 56 combinazioni rilevate in ASIC rientrano probabilmente in questa categoria.

È stata eseguita un'ulteriore procedura di filtraggio per valutare le coppie di parole estratte, consistente in un esperimento di *crowd sourcing* con parlanti madrelingua italiani. Ciò ha portato a una ulteriore selezione manuale di quelle combinazioni giudicate dai partecipanti come collocate. La scelta di questo tipo di esperimento consente al ricercatore di limitare un intervento aprioristico e di validare i dati ottenuti nel modo più oggettivo possibile.

Le procedure di *crowd sourcing* sono state tradizionalmente utilizzate in linguistica in vari campi, come la traduzione (Callison-Burch, 2009) e lo sviluppo di corpora (Wang *et al.*, 2012). L'originalità di questo studio si basa sul fatto che i parlanti nativi non sono impegnati in un processo di raccolta o definizione di dati, ma nella validazione degli stessi.

Ai partecipanti viene richiesto di svolgere un compito decisionale valutando se una data espressione può essere considerata una collocazione accademica italiana, in virtù della loro competenza metalinguistica e del loro retroterra educativo, poiché lo studio riguarda la lingua accademica. Interpretando la lingua come un processo sociale e condiviso, i parlanti nativi, che riconoscono intuitivamente la corretta composizione, il significato e l'uso delle collocazioni della loro lingua materna, sono chiamati a selezionare i dati e a stabilire cosa includere o meno in elenco. Una collocazione può quindi essere definita, in questo senso, come una particolare combinazione di parole che viene interpretata e valutata dai parlanti nativi grazie a una cornice esperienziale comune condivisa (Peppoloni, 2018b). Il significato di ogni combinazione, inteso come la somma dei fattori di accettabilità, comprensione e uso, deriva dal comune accordo che gli ipotetici attori di un evento comunicativo stabiliscono, grazie alla loro precedente esperienza nell'uso di

una data lingua in diverse situazioni comunicative. Ciò che distingue una collocazione da una combinazione libera, quindi, è il criterio di selezione attivato quando si utilizza una determinata combinazione di due o più parole per trasmettere un significato specifico che altre combinazioni lessicali non veicolerebbero. In altre parole, il fatto che *principio generale* sia considerato una collocazione deriva dal fatto che, all'interno di un quadro esperienziale coerente e condiviso tra i parlanti, esso assume un significato specifico. Allo stesso modo, *cosa diversa* è considerata una combinazione libera, poiché non viene riconosciuta come una combinazione di parole consolidata dall'uso in determinati contesti comunicativi. I soggetti coinvolti nell'esperimento sono stati 11 docenti universitari, parlanti nativi dell'italiano, di diversa età (5 compresi nella fascia di età tra i 30 e i 45 anni, 4 in quella tra i 46 e i 59 anni, e 2 in quella tra i 60 e i 70 anni) e di diverso genere (6 uomini e 5 donne), con una consuetudine d'uso della lingua accademica parlata e un'elevata competenza linguistica e metalinguistica, poiché l'indagine riguarda il lessico accademico e un particolare tipo di espressione polirematica.

Ogni partecipante ha ricevuto un modulo con un'introduzione agli obiettivi del progetto, le istruzioni necessarie su come prendere parte all'esperimento e un elenco di cinque o sei collocazioni da valutare. Nel caso in cui una data combinazione lessicale venga giudicata come una collocazione della lingua italiana, l'informante è invitato a provare a definirla o a fornirne un esempio. Di seguito si riporta un esempio delle brevi definizioni fornite da alcuni tra i partecipanti:

(1) **senso stretto: Sì**

Espressione di derivazione latina, (in) senso stretto si usa per descrivere qualcosa la cui definizione aderisce letteralmente al suo significato originario (smontando a volte l'uso metaforico di determinate espressioni) o ne restringe l'uso alla sua applicazione più immediata; es. Michele considera fratelli non solo la famiglia in senso stretto, ma anche i suoi amici di infanzia.

sistema operativo: Sì

Programma informatico di base necessario per accedere a tutte le funzioni di un computer: es. il sistema operativo è andato in crash, l'azienda ha prodotto un nuovo sistema operativo.

Interessanti sono alcune definizioni elaborate a partire dal riferimento ad altre combinazioni possibili, affini semanticamente, ma non esattamente sovrapponibili nei diversi contesti d'uso. Questo tipo di definizione rende evidente la riflessione metalinguistica effettuata dai parlanti per eseguire il compito richiesto e il livello di competenza linguistica richiesto:

(2) **ente pubblico: Sì**

Alternative: amministrazione pubblica; ente locale; ufficio pubblico (**non sono proprio sinonimi**) oppure Comune, Provincia, Regione, Ministero, etc. Es. La scuola è un ente pubblico essenziale per la vita dello Stato.

energia elettrica: Sì

Esempio: Oggi molte auto sono alimentate a energia elettrica
In molti casi è possibile usare l'espressione corrente elettrica, ma le due espressioni non sono perfettamente sovrapponibili.

Infine, un altro tipo di riflessione interessante compiuta dai partecipanti è quella relativa a coppie di parole che vengono giudicate come collocazioni della lingua italiana in

associazione ad altri elementi linguistici, a comporre delle locuzioni. Tale indicazione riflette la capacità dei parlanti nativi di analizzare la lingua per blocchi di elementi e non in maniera puramente sequenziale, anche in assenza del contesto d'uso di tali espressioni. Quando vengono esposti a una parte dell'espressione, i parlanti riescono in autonomia a richiamare i restanti elementi che la completerebbero:

(3) tempo stesso: Sì

Locuzione “al tempo stesso”, es. La situazione era imbarazzante e, al tempo stesso, preoccupante.

senso stretto: Sì

Di derivazione latina, (in) senso stretto si usa per descrivere qualcosa la cui definizione aderisce letteralmente al suo significato originario (smontando a volte l'uso metaforico di determinate espressioni) o ne restringe l'uso alla sua applicazione più immediata; es. “Michele considera fratelli non solo la famiglia in senso stretto, ma anche i suoi amici di infanzia”.

Gli 11 partecipanti a questa fase dello studio hanno identificato come collocazioni della lingua italiana 29 delle 56 disponibili, ovvero il 52% delle stesse⁷. Alcune vengono definite come di *uso comune*, come nel caso di *vita quotidiana* o *modo diverso*, per altre viene identificato un ambito semantico prevalente di utilizzo come nel caso di *intelligenza artificiale*, dove si parla di lessico dell'informatica, o di *corpo umano*, dove si fornisce un esempio che è tipico dei manuali di biologia, ovvero *il corpo umano è composto per il 60% di acqua*. In un caso viene evidenziata indirettamente la specificità d'uso delle collocazioni, identificandone il contesto di applicazione prevalente; è il caso di *attività professionale*, espressione a cui l'informante attribuisce il valore di collocazione, ma, sostiene, solo all'interno dei curricula professionali.

Nel caso di combinazioni che entrano a far parte di locuzioni più ampie, i parlanti sembrano non mostrare dubbi nell'identificarne il valore collocazionale; è il caso di combinazioni come *(al) tempo stesso*, *(in) senso stretto*, *(a) occhio nudo*, *(a) costo zero*. Inoltre, i parlanti sono sempre capaci di fornire degli esempi d'uso di tali espressioni, dimostrando di gestire i contesti d'uso della loro L1 con un elevato grado di competenza.

Ciò è evidente anche nell'espressione *occhio aperto*; sebbene questa sia fornita al singolare, solitamente viene utilizzata al plurale. Tale anomalia non sembra confondere gli informanti, che riconoscono la collocazione, la definiscono e citano degli esempi d'uso strutturati alla forma plurale:

(4) occhio aperto: Sì

L'espressione, utilizzata al plurale, è diffusa nella fraseologia italiana, ad indicare ad esempio uno stato di allerta (“Tieni gli occhi aperti!”) o uno stato di coscienza (“tu sogni ad occhi aperti”).

Utilizzata anche nell'espressione “sognare a occhi aperti”, divagare con l'immaginazione. Esempio: Sei distratto? Smettila di sognare a occhi aperti.

Dalle definizioni degli apprendenti si evince anche la loro capacità di cogliere il significato compositivo delle combinazioni riconosciute come collocazioni, non assimilabile alla somma dei singoli elementi che le compongono. Da questo emerge la capacità dei parlanti nativi di rapportarsi agilmente con i significati fraseologici della loro

⁷ Nell'Appendice 3 è possibile visualizzare l'elenco completo delle collocazioni identificate come tali e definite o esemplificate dai partecipanti.

L1, definiti in questa sede dagli informanti come *metaforici*, intuendo la specificità semantica rispetto ai singoli significati delle parole che entrano in combinazione:

(5) costo zero: Sì

È una espressione per indicare l'assenza di costo o metaforicamente di sforzo e di azione per ottenere un certo obiettivo (non puoi migliorare la tua vita a costo zero / alla fine abbiamo avuto l'auto nuova a costo zero).

Tale valore metaforico sembra uno dei fattori discriminanti nell'identificare le combinazioni proposte come collocazioni, tanto che l'espressione *anno precedente* non viene considerata una collocazione della lingua italiana, mentre al contrario *anno nuovo* viene validata come tale:

(6) anno nuovo: Sì

È una espressione per indicare l'inizio dell'anno (con l'anno nuovo siamo stati più fortunati) oppure l'anno successivo (nell'anno nuovo saremo più fortunati).

Come emerge dagli esempi forniti, l'espressione *anno nuovo* non identifica solo letteralmente un lasso di tempo che deve ancora attuarsi, ma anche un tempo del passato di cui si vuole enfatizzare l'inizio.

La frequenza delle collocazioni analizzate dagli informanti non sembra influire sulla loro capacità di validazione, difatti appartengono all'insieme delle espressioni selezionate sia quelle ad alta frequenza come *sistema operativo* (freq. 32) che quelle a bassa frequenza come *studente universitario* (freq. 3). Quest'ultima combinazione lessicale viene giudicata una collocazione proprio perché tipica della comunicazione accademica; lo stesso esempio che viene riportato dall'informante fa riferimento al comunicato di una biblioteca, luogo di studio privilegiato per questa categoria di utenti e, quindi, altamente contestuale (i.e. *il prestito dei volumi è riservato agli studenti universitari*).

Tutte le collocazioni selezionate dai partecipanti allo studio sono state ulteriormente confrontate con l'elenco disponibile nel *Dizionario delle Collocazioni* di Tiberii (2012), al fine di avere un processo bilanciato di doppia validazione dei dati, basato da un lato sulla competenza, anche di carattere impressionistico, dei parlanti nativi e, dall'altro, sulla sistematicità di un'opera lessicografica. Il suddetto dizionario comprende 6.000 voci tra aggettivi, nomi e verbi. Ogni lemma è presentato in associazione a una selezione delle collocazioni più significative in cui è coinvolto, chiaramente organizzate sulla base del loro uso e delle relative categorie grammaticali di appartenenza. L'autrice ha scelto di eliminare, ai due estremi, sia le collocazioni rare o eccessivamente specifiche, che quelle troppo generali e onnicomprensive. Non compaiono come voci, quindi, aggettivi come *bello*, *brutto*, *buono*, *vecchio* e verbi come *avere* ed *essere*, a meno che questi elementi, in combinazione con altre parole, non diano origine a collocazioni facilmente riconoscibili (e.g. *buon appetito*, *avere buongusto*, *essere in cammino*, ecc.)

Delle 29 combinazioni lessicali identificate dai partecipanti come collocazioni, solo 7 non trovano spazio nell'opera di Tiberii, ovvero il 24%. In alcuni casi si tratta di quelle espressioni che fanno parte di locuzioni più ampie, come (*a*) *occhio nudo*, altre sono di carattere troppo generale, come *modo migliore* o *tempo stesso* e, per questo, non incontrano i parametri costitutivi dell'opera.

In definitiva, la competenza metalinguistica dei parlanti nativi potrebbe essere utilizzata sia per raffinare le liste di coppie di parole estratte automaticamente da un corpus, sia per ampliare la gamma di collocazioni da includere in una ipotetica risorsa lessicografica.

5. CONCLUSIONI

I risultati dello studio mostrano come le collocazioni possano costituire un elemento complesso della lingua sia per i sistemi automatici che per i parlanti. Ciò che emerge da un punto di vista computazionale è che i metodi *P-based* mostrano una precisione maggiore con collocazioni relativamente fisse, brevi e adiacenti. I metodi a dipendenza sintattica potrebbero invece essere utili per rilevare collocazioni discontinue e sintatticamente flessibili, perché capaci di identificare parole sintatticamente correlate indipendentemente dalle loro realizzazioni superficiali. Le collocazioni N-Adj non mostrano un alto grado di inserzione di elementi linguistici terzi, se non per ciò che riguarda gli avverbi, come nel caso di *modo molto migliore*, ma se in futuro si volesse lavorare su altre tipologie collocazionali, come quelle V-N, allora un ulteriore passaggio della nostra ricerca potrebbe essere quello di analizzare sintatticamente il corpus ASIC, integrando le informazioni derivanti dalle procedure di *POS-tagging*.

Come già specificato nella sezione dedicata alla descrizione dell'estrazione e dell'analisi dei dati, la frequenza, di per sé, non sembra essere un parametro efficace nella gestione della collocatività di un'espressione. Per raffinare questo parametro, in futuro, si potrebbe calcolare il grado di attrazione tra i componenti di una data combinazione tramite l'uso aggiuntivo di misure associative, quali la *Mutual Information*. Uno studio precedente di Evert (2008), riguardante la lingua inglese, metteva a confronto la produzione di collocazioni da parte di parlanti nativi e quella di apprendenti dell'inglese. I risultati sembrano indicare che i parlanti nativi producono combinazioni lessicali che mostrano un alto grado di associazione, identificabile tramite la misura associativa della *Mutual Information*, piuttosto che un'alta frequenza nel corpus di riferimento, come accade per gli apprendenti stranieri. Di conseguenza, le collocazioni dei parlanti nativi sembrano essere più salienti che frequenti. La caratteristica della salienza conferisce evidente idiomaticità alle produzioni dei parlanti nativi e, al contrario, la loro moderata presenza nelle produzioni degli apprendenti sarebbe la ragione per cui le loro collocazioni sembrano perdere naturalezza ed efficacia comunicativa.

La salienza e il forte grado di associazione sembrano quindi essere fattori chiave nell'identificazione delle collocazioni tipiche dei parlanti nativi. Poiché il corpus ASIC contiene solo testi prodotti da parlanti nativi, la frequenza delle combinazioni N-Adj potrebbe probabilmente essere un primo filtro statistico nei dati, ulteriormente integrato da una misura associativa che rilevi il loro grado di attrazione e quindi di idiomaticità.

I risultati dell'esperimento di *crowd sourcing* sembrano indicare che i parlanti nativi hanno sia un'idea intuitiva di cosa siano le collocazioni e di come vengano utilizzate nella loro lingua materna nel contesto accademico, sia la capacità di riflettere metalinguisticamente, in modo esplicito, su di esse, identificandole, definendone l'uso e il significato. Il compito loro proposto in questo studio era di natura ancora più complessa, poiché i partecipanti potevano fare riferimento solo alle coppie di parole proposte e non ai loro contesti linguistici effettivi d'uso contenuti nel corpus, trovandosi quindi a dover elaborare per proprio conto esempi d'uso reale. Sebbene l'uso delle collocazioni faccia parte della competenza metalinguistica di un parlante nativo (Gombert, 1990), esso è comunque strettamente legato alla componente contestuale, che avrebbe la capacità di innescare nei parlanti delle riflessioni esplicite su tali combinazioni lessicali. Probabilmente, la capacità dei soggetti coinvolti di identificare quelle che ritenevano essere collocazioni all'interno dell'elenco fornito e di produrre una definizione e degli esempi d'uso è fortemente correlata alla familiarità degli stessi con il contesto comunicativo accademico e all'elevata capacità di analisi e controllo della lingua (Bialystock, 2001; Peppoloni, 2018a), come dimostrato dal loro titolo di studio e dalla professione svolta.

Un altro punto rilevante è che molte delle collocazioni identificate potrebbero provenire dal linguaggio generale e non specificamente dal dominio accademico, poiché quest'ultimo è un sottogruppo del primo (Nation, 2001). Una possibile soluzione potrebbe essere l'introduzione, oltre all'indice di frequenza, di un altro parametro statistico, quello della *keyness*. Questa misura supporta gli utenti nell'identificazione delle parole chiave in una raccolta di testi, vale a dire i lemmi più altamente caratterizzanti, dimostrando una frequenza non comune per una determinata varietà linguistica, come quella accademica. Inoltre, dato che la lingua accademica viene utilizzata per argomentare su diversi saperi disciplinari, l'altro rischio, all'opposto, è quello di estrarre collocazioni settoriali, cioè non trasversali ai diversi contenuti disciplinari, ma tipici di un certo contesto di specializzazione. In questo studio si è cercato di ovviare a questo problema, filtrando le collocazioni sulla base della loro distribuzione all'interno del corpus ASIC, ovvero includendo nell'elenco sottoposto alla validazione dei parlanti nativi solo quelle combinazioni lessicali che fossero presenti in almeno due dei settori del sapere inclusi in ASIC.

Sulla base dei risultati di questo studio preliminare, possiamo delineare altri sviluppi futuri attesi. In primo luogo, la dimensione del corpus potrebbe essere ampliata. Questo cambiamento potrebbe essere utile per eliminare il problema computazionale della scarsità dei dati. Un altro possibile sviluppo potrebbe derivare dall'aumento del campione di soggetti coinvolti nell'esperimento di *crowd sourcing* e nella diversificazione delle loro caratteristiche idiosincratice, al fine di garantire un campione più rappresentativo della popolazione di informanti. Infine, sarebbe utile elaborare liste di collocazione multilingui, poiché questi fenomeni linguistici costituiscono un problema acquisizionale cruciale, in quanto altamente convenzionali, arbitrari e privi di corrispondenza interlinguistica. La produzione di strumenti e risorse utili a sviluppare percorsi didattici incentrati sui fenomeni fraseologici della lingua, favorirebbero l'esposizione strutturale e contestuale degli apprendenti stranieri, migliorando il loro livello di competenza linguistico-comunicativa e inducendoli a un confronto metalinguistico con la propria L1 e con le altre lingue eventualmente appartenenti al loro repertorio.

RIFERIMENTI BIBLIOGRAFICI

- Anstrom K., Dicerbo P., Butler F., Katz A., Millet J., Rivera C. (2010), *A review of the literature on academic English: Implications for K-12 English language learners*, The George Washington University Center for Equity and Excellence in Education, Arlington.
- Bailey A. L., Heritage H. M. (2008), *Formative assessment for literacy, grades K-6: Building reading and academic language skills across the curriculum*, Corwin Press, Thousand Oaks.
- Bialystok E. (2001), *Bilingualism in Development: Language, Literacy, and Cognition*, Cambridge University Press, Cambridge.
- Biber D. (2006), *University Language: a corpus-based study of spoken and written registers*, John Benjamins, Amsterdam.
- Biber D., Conrad S., Rappen R. (1998), *Corpus linguistics: Exploring language structure and use*, Cambridge University Press, Cambridge.
- Bourdieu P., Passeron J. C. (1994), "Introduction: Language and relationship to language in the teaching situation", in Bourdieu P., Passeron J. C., de Saint Martin M. (eds.), *Academic Discourse: Linguistic misunderstanding and professorial power*, Polity, Cambridge, pp. 1-34.

- Callison-Burch C. (2009), "Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk", in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 286-295.
- Corpas Pastor G. (2003), *Diez años de investigaciones en fraseología: análisis sintáctico-semánticos, contrastivos y traductológicos*, Iberoamericana-Vervuet, Madrid.
- Cowie A. P. (1988), "Stable and creative aspects of vocabulary use", in Carter R., McCarthy M. J. (eds.), *Vocabulary and Language Teaching*, Longman, London, pp. 126-137.
- Cowie A. P. (1992), "Multiword Lexical Units and Communicative Language Teaching", in Arnaud P. J. L., Béjoint H. (eds.), *Vocabulary and Applied Linguistics*, Palgrave Macmillan, London, pp. 1-12.
- Coxhead A. (2000), "A New Academic Word List", in *TESOL Quarterly*, 34, pp. 213-38.
- Evert S. (2008), "Corpora and collocations", in Lüdeling A, Kytö M (eds.), *Corpus Linguistics. An International Handbook*, Mouton de Gruyter, Berlin, New York, pp. 1212-1248.
- Facione P. (1990), *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction*, California Academic Press, Millbrae.
- Farrell P. (1990), *Vocabulary in ESP: A lexical analysis of the English of electronics and a study of semi-technical vocabulary* (CLCS Occasional Paper No. 25), Trinity College, Centre for Language and Communication Studies.
- Fernando C. (1996), *Idioms and idiomaticity*, Oxford University Press, Oxford.
- Fillmore C., Kay P., O'Connor M. C. (1988), "Regularity and idiomaticity in grammatical constructions: the case of let alone", in *Language*, 64, pp. 501-538.
- Firth J. R. (1957), "Modes of Meaning", in Palmer F. R. (ed.), *Papers in Linguistics 1934-1951*, Oxford University Press, Oxford, pp. 190-215.
- Gombert J. E. (1990), *Le développement métalinguistique*, PUF, Paris.
- Grace G. (1987), *The Linguistic Construction of Reality*, Croom Helm, New York.
- Hamp-Lyons L. (2011), "English for Academic Purposes", in Hinkel E. (ed.), *Handbook of Research in Second Language Teaching and Learning*, Routledge, Abingdon, pp. 89-105.
- Hyland K. (2002), "Specificity revisited: How far should we go now?", in *English for Specific Purposes*, 21, 4, pp. 385-395.
- Ježek E., Sprugnoli R. (2023), *Linguistica computazionale*, il Mulino, Bologna.
- Jones S., Sinclair J. M. (1974), "English lexical collocations. A study in computational linguistics", in *Cahiers de Lexicologie*, 24, pp. 15-61.
- Jordan R. R. (1997), *English for academic purposes: a guide and resource book for teachers*, Cambridge University Press, Cambridge.
- Kilgarriff A., Rychly P., Smrz P., Tugwell D. (2004), "The Sketch Engine", in *Proceedings EURALEX*, Lorient, pp. 105-116.
- Krishnamurthy R. (2006), "Collocations", in *Encyclopedia of Language and Linguistics*, vol. 2, pp. 596-600.
- Laufer B. (2011), "The Contribution of Dictionary Use to the Production and Retention of Collocations in a Second Language", in *International Journal of Lexicography*, 24, 1, pp. 29-49.
- Marazzini C. (2021), "La lingua di Dante non può parlare di scienza. Il MUR esclude l'italiano nel bando per i fondi FIS", in *Italiano digitale*, 19, 4, pp. 181-183.
- Mcenery T., Hardie A. (2011), *Corpus Linguistics: Method, Theory and Practice*, Cambridge University Press, Cambridge.
- Nagy W., Townsend D. (2012), "Words as tools: Learning academic vocabulary as language acquisition", in *Reading Research Quarterly*, 47, pp. 91-108.

- Nation I. S. P. (2001), *Learning vocabulary in another language*, Cambridge University Press, Cambridge.
- Nattinger J. R., DeCarrico J. S. (1992), *Lexical Phrases and Language Teaching*, Oxford University Press, Oxford.
- Nesselhauf N. (2005), *Collocations in a learner corpus*, John Benjamins Publishing, Amsterdam.
- Pawley A., Syder F. H. (1983), “Two puzzles for linguistic theory: nativelike selection and nativelike fluency”, in Richards J., Schmidt R. (eds.), *Language and Communication*, Longman, London, pp. 191-226.
- Peppoloni D. (2018a), *Glottodidattica e metalinguaggio*, Guerra Editore, Perugia.
- Peppoloni D. (2018b), “A corpus-based study of the automatic extraction and validation of V-N Italian oral academic collocations”, in *Linguisticae Investigationes*, 41, 2, pp. 240-268.
- Schleppegrell M. J. (2004), *The language of schooling: A functional linguistics perspective*, Lawrence Erlbaum Associates Publishers.
- Sinclair J. (1991), *Corpus, Concordance, Collocation*, Oxford University Press, Oxford.
- Snow R., O'Connor B., Jurafsky D., Ng A. (2008), “Cheap and fast – but is it good?: evaluating non-expert annotations for natural language tasks”, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics*, pp. 254-263.
- Tiberii P. (2012), *Il dizionario delle collocazioni*, Zanichelli, Bologna.
- Vygotsky L. S. (1987), “Thinking and Speech”, in Rieber R. W., Carton A. S. (eds.), *The Collected Works of L. S. Vygotsky* (Vol. 1), *Problems of General Psychology*, Plenum Press, New York, pp. 39-285 (Original Work Published, 1934).
- Wang A., Hoang C., Kan M. Y. (2012), “Perspectives on Crowdsourcing Annotations for Natural Language Processing”, in *Language Resources and Evaluation*, 47, 1, pp. 9-31.
- Wray A. (2002), *Formulaic language and the lexicon*, Cambridge University Press, Cambridge.
- Wray A., Perkins M. (2000), “The functions of formulaic language: An integrated model”, in *Language & Communication*, 20, 1, pp. 1-28.
- Yorio C. (1989), “Idiomaticity as an indicator of second language proficiency”, in Hyltenstam K., Obler Lobler L. K. (eds.), *Bilingualism across the Lifespan*, Cambridge University Press, Cambridge, pp. 55-72.

APPENDICE 1.

Lista delle collocazioni estratte automaticamente con frequenza = o > di 3, con indicazione dell'area disciplinare di appartenenza

Lemma	Frequenza	Area discipl.	Area discipl.	Area discipl.	Area discipl.
interesse pubblico	60	Giurecon			
vertice politico	47	Giurecon			
potere amministrativo	39	Giurecon			
procedura remoto	38			Scient	
gruppo funzionale	35		Med		
organo politico	33	Giurecon			
sistema operativo	32	Giurecon		Scient	
mano sinistra	31		Med		

atto amministrativo	24	Giurecon			
indirizzo politico	24	Giurecon			
ente pubblico	23	giurecon			Uman
forma vera	21			Scient	
volta scorsa	20		Med	Scient	Uman
potere discrezionale	19	Giurecon			
servizio esclusivo	19	Giurecon			
atteggiamento implicito	18				Uman
procedimento amministrativo	17	Giurecon			
risorsa naturale	16				Uman
persona umana	16	Giurecon			
scuola superiore	15				Uman
atteggiamento esplicito	15				Uman
catena laterale	14		Med		
simbolo terminale	13			Scient	
interesse secondario	13	Giurecon			
debito pubblico	13	Giurecon			
decisione amministrativa	13	Giurecon			
tempo stesso	12	Giurecon	Med	Scient	Uman
chiusura riflessiva	12			Scient	
essere umano	12	Giurecon	Med		Uman
area associativa	12		Med		
modo diverso	12	Giurecon	Med		Uman
funzionario pubblico	12	Giurecon			
dirigente scolastico	12				Uman
interesse primario	12	Giurecon			
forma negata	11			Scient	
funzione amministrativa	11	Giurecon			
intelligenza artificiale	11	Giurecon			Uman
area secondaria	10		Med		
sviluppo sostenibile	10				Uman
interesse privato	10	Giurecon			
opera pubblico	10	Giurecon			
piano urbanistico	10	Giurecon			
decisione generale	10	Giurecon			
carattere creativo	10	Giurecon			
brano musicale	10	Giurecon	Med		
gruppo sociale	9				Uman
cosa diversa	9		Med	Scient	Uman
ordine economico	9	Giurecon			
atteggiamento linguistico	9				Uman
percorso logico	9	Giurecon	Med		
tragedia antica	8				Uman
testo unico	8	Giurecon			
scuola media	8				Uman
potere espressivo	8			Scient	

percorso decisionale	8	Giurecon			
funzione cognitiva	8		Med		
anno successivo	8		Med		Uman
informazione visiva	8		Med		
generazione futura	8				Uman
decisione pubblica	8	Giurecon			
struttura primaria	8		Med		
parte sinistra	8		Med		
procedura selettiva	8	Giurecon			
risorsa umana	7				Uman
identità sociale	7				Uman
centro storico	7	Giurecon			Uman
freccia gialla	7			Scient	
cultura umanistico	7				Uman
logica dinamica	7			Scient	
grammatica regolare	7			Scient	
area corrispondente	7		Med		
persona anziana	7		Med		
maniera diversa	7		Med	Scient	Uman
principio generale	7	Giurecon	Med		Uman
momento storico	7				Uman
interesse concreto	7	Giurecon			
attività economica	7	Giurecon			
scelta discrezionale	7	Giurecon			
secolo futuro	7				Uman
opera letteraria	7	Giurecon			
firma digitale	7	Giurecon			
attività amministrativa	7	Giurecon			
comunicazione scientifica	7	Giurecon			Uman
crisi economica	6	Giurecon			Uman
atteggiamento negativo	6				Uman
psicologia sociale	6				Uman
ambito accademico	6	Giurecon	Med		
amministrazione pubblico	6	Giurecon			Uman
tratto ottico	6		Med		
responsabilità ministeriale	6	Giurecon			
decisione concreta	6	Giurecon			
linguaggio vuoto	6			Scient	
potere decisionale	6	Giurecon			
cosa interessante	6		Med	Scient	
caso concreto	6	Giurecon			
maniera corretta	6	Giurecon	Med		Uman
scelta organizzativa	6	Giurecon			
mensa scolastica	6	Giurecon			
direttore generale	6	Giurecon			
proprietà intellettuale	6	Giurecon			

regola generale	6	Giurecon	Med		
parte superiore	6		Med	Scient	
senso stretto	5		Med		Uman
livello basso	5			Scient	
testo costituzionale	5	Giurecon			
reazione emotiva	5				Uman
superficie terrestre	5				Uman
congiuntura astrale	5				Uman
scelta linguistica	5				Uman
gruppo chiuso	5			Scient	
comunità parlante	5				Uman
linguaggio naturale	5			Scient	
comunità italiana	5				Uman
caso particolare	5			Scient	
stella fissa	5				Uman
programmazione parallela	5			Scient	
elemento fondamentale	5	Giurecon	Med		Uman
gruppo aperto	5			Scient	
attività lavorativa	5	Giurecon			
cosa nuova	5	Giurecon	Med		Uman
interesse esterno	5	Giurecon			
parte inferiore	5		Med		
risorsa economica	5				Uman
decisione finale	5	Giurecon			
patrimonio culturale	5	Giurecon			
posizione dominante	5	Giurecon			
vita quotidiana	5	Giurecon	Med		
occhio nudo	5		Med		Uman
interesse principale	5	Giurecon			
relazione causale	5			Scient	
modo migliore	5	Giurecon	Med	Scient	
funzione pubblica	5	Giurecon			Uman
struttura secondaria	5		Med		
chilometro zero	5				Uman
forma normale	5			Scient	
potenza assoluta	4				Uman
sostenibilità debole	4				Uman
energia elettrica	4	Giurecon			Uman
modo imparziale	4	Giurecon			
dirigente generale	4	Giurecon			
essere vivente	4		Med		
anno nuovo	4		Med		Uman
ordinamento globale	4			Scient	
momento diverso	4		Med	Scient	Uman
tempo totale	4			Scient	
freccia arancione	4			Scient	

impiegato pubblico	4	Giurecon			
derivazione diretta	4			Scient	
problema legato	4		Med	Scient	
docente universitario	4				Uman
storia personale	4		Med		
principio generale	4	Giurecon	Med		Uman
livello diverso	4	Giurecon			Uman
risposta emotiva	4		Med		Uman
relazione lavorativa	4	Giurecon			
interesse sociale	4	Giurecon			Uman
interesse diffuso	4	Giurecon			
anno passato	4		Med	Scient	Uman
capitale naturale	4				Uman
responsabilità proprie	4	Giurecon			
versante oggettivo	4	Giurecon			
elemento importante	4	Giurecon			Uman
caso specifico	4	Giurecon	Med	Scient	
conoscenza scientifica	4	Giurecon			Uman
fine ultimo	4	Giurecon			Uman
esame orale	4		Med		
natura pubblica	4	Giurecon			
relazione sociale	4		Med		
decisione puntuale	4	Giurecon			
misura organizzativa	4	Giurecon			
parco nazionale	4				Uman
risposta istintiva	4		Med		
competenze proprie	4	Giurecon			
linea generale	4		Med		
previsione costituzionale	4	Giurecon			
maniera efficace	4	Giurecon			Uman
articolo scientifico	4	Giurecon			
data certa	4	Giurecon			
uomo adulto	4		Med		
direttiva generale	4	Giurecon			
concetto generale	4		Med		
ambiente esterno	4		Med		
anno precedente	4		Med		Uman
espressione linguistica	4				Uman
autorità amministrativa	3	Giurecon			
scienza moderna	3				Uman
lezione successiva	3				Uman
scala numerica	3				Uman
elemento neutro	3			Scient	
conflitto strutturale	3	Giurecon			
risorsa rinnovabile	3				Uman
indice sommario	3	Giurecon			

corpo umano	3		Med		Uman
discorso diverso	3				Uman
risorsa alimentare	3				Uman
grazia divina	3				Uman
contesto sociale	3				Uman
mondo sensibile	3				Uman
vista logico	3				Uman
atto creativo	3	Giurecon			
maniera efficiente	3	Giurecon			
estensione massima	3				Uman
lingua straniera	3				Uman
potere predittivo	3				Uman
determinazione specifica	3				Uman
maniera uniforme	3	Giurecon			
dinamica sociale	3				Uman
riflessione filosofica	3				Uman
questione filosofica	3				Uman
contesto educativo	3				Uman
differenza sostanziale	3		Med		
tema fondamentale	3				Uman
genere umano	3				Uman
vita professionale	3		Med		
situazione soggettiva	3	Giurecon			
fenomeni corruttivi	3	Giurecon			
forma canonica	3			Scient	Uman
evento storico	3				Uman
carattere generale	3	Giurecon			
processo decisionale	3	Giurecon			
livello applicativo	3			Scient	
processo preparativo	3		Med		
potere generativo	3			Scient	
processo generativo	3			Scient	
volta celeste	3				Uman
protocollo sperimentale	3	Giurecon			
persone adulte	3	Giurecon			
condizione stazionaria	3			Scient	
profili professionali	3	Giurecon			
situazione diversa	3	Giurecon			Uman
effetto dannoso	3	Giurecon			
cosa importante	3	Giurecon		Scient	Uman
potere gerarchico	3	Giurecon			
maniera coerente	3	Giurecon			
avviso pubblico	3	Giurecon			
occhio aperto	3		Med		Uman
lavoratore pubblico	3	Giurecon			
termine perentorio	3	Giurecon			

partito politico	3	Giurecon			
zona grigia	3		Med		
punto interrogativo	3		Med		Uman
metodo scientifico	3				Uman
attività culturale	3	Giurecon			
servizio pubblico	3	Giurecon			
organizzazione interna	3	Giurecon			
opinione pubblica	3	Giurecon			Uman
analisi formale	3				Uman
bestia feroce	3				Uman
modello organizzativo	3	Giurecon			
valore artistico	3	Giurecon			
dirigenza pubblica	3	Giurecon			
scelta politica	3	Giurecon			
classe energetica	3	Giurecon			
organo competente	3	Giurecon			
figura professionale	3				Uman
attività professionale	3	Giurecon			Uman
spesa pubblica	3	Giurecon			
costo zero	3	Giurecon		Scient	
concetto matematico	3	Giurecon			
studente universitario	3		Med		Uman
centro urbano	3	Giurecon			Uman
livelli essenziali	3	Giurecon			
lezione precedente	3		Med	Scient	
casa editrice	3	Giur econ			
concetto fondamentale	3		Med	Scient	Uman
excursus storico	3	Giurecon			Uman
faccia felice	3		Med		
lezione teorica	3		Med		
insieme infinito	3			Scient	
articolo giornalistico	3				Uman
scienza ufficiale	3		Med		
caso clinico	3		Med		

APPENDICE 2.

Lista delle 56 collocazioni distribuite in almeno due delle aree disciplinari contenute in ASIC

Lemma	Frequenza	Area discipl.	Area discipl.	Area discipl.	Area discipl.
1. sistema operativo	32	Giurecon		Scient	
2. ente pubblico	23	Giurecon			Uman
3. volta scorsa	20		Med	Scient	Uman
4. tempo stesso	12	Giurecon	Med	Scient	Uman
5. essere umano	12	Giurecon	Med		Uman

6. modo diverso	12	Giurecon	Med		Uman
7. intelligenza artificiale	11	Giurecon			Uman
8. brano musicale	10	Giurecon	Med		
9. cosa diversa	9		Med	Scient	Uman
10. percorso logico	9	Giurecon	Med		
11. anno successivo	8		Med		Uman
12. centro storico	7	Giurecon			Uman
13. maniera diversa	7		Med	Scient	Uman
14. principio generale	7	Giurecon	Med		Uman
15. comunicazione scientifica	7	Giurecon			Uman
16. crisi economica	6	Giurecon			Uman
17. ambito accademico	6	Giurecon	Med		
18. osa interessante	6		Med	Scient	
19. maniera corretta	6	Giurecon	Med		Uman
20. regola generale	6	Giurecon	Med		
21. parte superiore	6		Med	Scient	
22. senso stretto	5		Med		Uman
23. elemento fondamentale	5	Giurecon	Med		Uman
24. cosa nuova	5	Giurecon	Med		Uman
25. vita quotidiana	5	Giurecon	Med		
26. occhio nudo	5		Med		Uman
27. modo migliore	5	Giurecon	Med	Scient	
28. funzione pubblica	5	Giurecon			Uman
29. energia elettrica	4	Giurecon			Uman
30. anno nuovo	4		Med		Uman
31. momento diverso	4		Med	Scient	Uman
32. problema legato	4		Med	Scient	
33. principio generale	4	Giurecon	Med		Uman
34. livello diverso	4	Giurecon			Uman
35. risposta emotiva	4		Med		Uman
36. interesse sociale	4	Giurecon			Uman
37. anno passato	4		Med	Scient	Uman
38. elemento importante	4	Giurecon			Uman
39. caso specifico	4	Giurecon	Med	Scient	
40. conoscenza scientifica	4	Giurecon			Uman
41. fine ultimo	4	Giurecon			Uman
42. maniera efficace	4	Giurecon			Uman
43. anno precedente	4		Med		Uman
44. corpo umano	3		Med		Uman
45. forma canonica	3			Scient	Uman
46. situazione diversa	3	Giurecon			Uman
47. cosa importante	3	Giurecon		Scient	Uman
48. occhio aperto	3		Med		Uman
49. punto interrogativo	3		Med		Uman
50. attività professionale	3	Giurecon			Uman
51. costo zero	3	Giurecon		Scient	

52. studente universitario	3		Med		Uman
53. centro urbano	3	Giurecon			Uman
54. lezione precedente	3		Med	Scient	
55. concetto fondamentale	3		Med	Scient	Uman
56. excursus storico	3	Giurecon			Uman

APPENDICE 3.

Elenco delle 29 combinazioni lessicali giudicate dai parlanti nativi come collocazioni

1. sistema operativo
2. ente pubblico
3. tempo stesso
4. essere umano
5. modo diverso
6. intelligenza artificiale
7. centro storico
8. maniera diversa
9. principio generale
10. comunicazione scientifica
11. crisi economica
12. regola generale
13. senso stretto
14. elemento fondamentale
15. vita quotidiana
16. occhio nudo
17. modo migliore
18. energia elettrica
19. anno nuovo
20. principio generale
21. risposta emotiva
22. caso specifico
23. fine ultimo
24. corpo umano
25. occhio aperto
26. attività professionale
27. costo zero
28. studente universitario
29. excursus storico

