COLL-IT: UNO STRUMENTO DI VALUTAZIONE DELLA COMPETENZA COLLOCAZIONALE. REALIZZAZIONE E SOMMINISTRAZIONE DI UN TEST SULLE COLLOCAZIONI VERBO-NOME PER APPRENDENTI DI ITALIANO L2/LS

Francesca La Russa[1], *Fabio Zanda*[2], *Anna Suadoni*[3]

1. INTRODUZIONE

Il lessico è una componente fondamentale nello sviluppo della competenza linguistico-comunicativa in lingua seconda (L2) e diversi sono i lavori condotti sulla valutazione di questa dimensione (sul piano internazionale si vedano ad esempio Read, 2000; Daller *et al.*, 2007; Milton, 2009; Schmitt, 2010; Nation, Webb, 2011; Bardel *et al.*, 2013; Dóczi, Kormos, 2016; Meara, Miralpeix, 2017; mentre per quel che riguarda l'italiano L2 si menziona il recente lavoro di Gallina, 2022).

La competenza lessicale include la conoscenza della forma e del significato di una parola, la capacità di comprenderla e produrla in determinati contesti di uso, ma anche la capacità di combinare tra loro due o più parole che occorrono frequentemente insieme. Come sottolinea Wray (2002: 143) infatti: «To know a language you must know not only its individual words, but also how they fit together». Quest'ultimo aspetto rimanda alla competenza collocazionale che implica la conoscenza dei rapporti associativi tra le parole (Nation, 2001) e, quindi, delle collocazioni.

Sono state date numerose definizioni del termine *collocazione*. Una definizione "ampia", di natura frequentista, comprende qualsiasi combinazione frequente di parole (Sinclair, 1991), una definizione più "stretta" include le combinazioni di parole soggette a una restrizione lessicale (*restricted lexical co-occurrence*, cfr. Mel'cuk, Wanner, 1994). Si riprende qui la definizione proposta direttamente per la lingua italiana da Jezek (2005: 192) secondo cui la collocazione è «una combinazione di parole consolidata dall'uso, corrispondente a un modo preferenziale di dire una certa cosa». -Tale combinazione di parole è «soggetta a una restrizione lessicale, per cui la scelta di una specifica parola (il collocato) per esprimere un determinato significato, è condizionata da una seconda parola (la base) alla quale questo significato è riferito». Fra gli esempi citati dall'autrice ci sono *pioggia battente* (in cui *pioggia* è la base e *battente* il collocato) o *stendere un documento* (*stendere* base e *documento* collocato).

In Spina (2016: 225), vengono presentati i tipi di combinazioni lessicali più frequenti estratte dal Perugia Corpus (PEC):

[1] Sapienza Università di Roma.
[2] Università per Stranieri di Perugia.
[3] Universidad de Granada.
Il lavoro nasce da una stretta collaborazione tra gli autori nell'ambito del progetto Prin PHRAME (20178XXKFY) - *Phraseological Complexity Measures in learner Italian*. Tuttavia, si devono a Francesca La Russa i paragrafi 1 e 3; a Fabio Zanda i paragrafi 5, 6 e 7; a Anna Suadoni i paragrafi 2 e 4.

verbo + (articolo) + nome (35%): *fare la doccia*, *avere paura*.
nome + preposizione + nome (34%): *opera d'arte*.
nome + aggettivo (17%): *sistema operativo*.
aggettivo + nome (5%): *terzo mondo*.
verbo + aggettivo (4%): *stare zitto*.
nome + nome (3%): *conferenza stampa*.
aggettivo + congiunzione + aggettivo (1%): *bianco e nero*.
nome + congiunzione + nome (1%): *andata e ritorno*[4].

Le combinazioni lessicali, tra cui le collocazioni, sono centrali nell'apprendimento linguistico. Studi psicolinguistici mostrano che vengono elaborate più velocemente rispetto alle sequenze inedite (Siyanova-Chanturia, 2015). Esse rappresentano inoltre una sorta di "isole di affidabilità" (*islands of reliability*, cfr. Henriksen, 2013) su cui gli apprendenti possono fare affidamento in produzione e in ricezione invece di dover costruire il messaggio parola per parola. L'uso di combinazioni fraseologiche da parte degli apprendenti permette quindi di ridurre lo sforzo di elaborazione e allo stesso tempo di incrementare la fluenza (Nesselhauf, 2005).

Recenti studi basati sulla linguistica dei corpora hanno permesso di osservare lo sviluppo della competenza collocazionale in apprendenti L2. Per l'italiano, citiamo in particolare i lavori di Siyanova-Chanturia (2015), Siyanova-Chanturia, Spina (2019), Omidian *et al.* (2020), Spina (2018; 2019; 2022). Tale sviluppo si è rivelato lento (Yoon, 2016), non lineare (Bestgen, Granger, 2014; Siyanova-Chanturia, Spina, 2019) e difficoltoso (Wang, 2016).

Vista l'utilità ma anche la difficoltà di sviluppare la competenza collocazionale, l'insegnamento delle collocazioni dovrebbe essere una priorità nei corsi di lingua (Lewis, 2000). Tuttavia, a differenza di altre unità fraseologiche come le frasi idiomatiche e i proverbi, le collocazioni non vengono sufficientemente messe in evidenza nell'insegnamento dell'Italiano L2 (Bini *et al.*, 2007; Berišić Antić, 2016). Similmente, se numerosi test sono stati realizzati per la valutazione di altri aspetti della competenza lessicale, quelli per la valutazione della competenza collocazionale in italiano L2/LS sono invece, almeno a nostra conoscenza, relativamente scarsi[5].

Il presente contributo si propone di colmare questa lacuna. Dopo una breve rassegna sui principali test di valutazione della competenza collocazionale per apprendenti di inglese, verrà infatti presentato il COLL-IT, un test sulle collocazioni verbo + (articolo) + nome (oggetto) per apprendenti di italiano L2 di livello B1, B2 e C del *Quadro Comune Europeo di Riferimento* (QCER; Consiglio d'Europa, 2002), realizzato a partire da un sillabo delle collocazioni italiane (La Russa *et al.*, 2023), e saranno discussi gli esiti della sua somministrazione a 103 apprendenti ispanofoni.

2. VALUTARE LA COMPETENZA COLLOCAZIONALE

Fra i modelli più citati per la descrizione della competenza lessicale c'è quello di Nation (2001: 27) che individua alcuni elementi determinanti per la conoscenza, ricettiva e produttiva, di una parola:

[4] A queste si aggiungono le combinazioni: *aggettivo + avverbio + nome'* e *verbo + avverbio* (Spina, 2016: 30).
[5] Si veda ad esempio il test sviluppato da Suadoni (2020) sulle collocazioni verbo-nome; quello di Pallone, (2023) sulle costruzioni a verbo supporto; il test di Gallina (2018) sul vocabolario accademico in cui compare una sezione dedicata alla valutazione delle collocazioni estratte dall'*Academic Italian Word List* (AIWL, Spina, 2010).

- form (spoken form, written form, word parts);
- meaning (form and meaning, concepts and referents, associations);
- use (grammatical functions, collocations, constraints on use).

Secondo Nation, per conoscere una parola è necessario individuarla quando viene pronunciata e scritta e saperla pronunciare e scrivere, interpretando correttamente fonemi e grafemi che possono variare nella flessione o per l'aggiunta di affissi. Di conseguenza, la conoscenza di una parola implica anche saper gestire e costruire le strutture grammaticali adeguate ad accoglierla ed essere in grado di prevedere che parole o che tipo di parole potrebbero precederla o seguirla. A questo si aggiunge che il parlante deve sapere quanto una parola è frequente e in quali contesti è appropriata. Infine, deve conoscerne il significato, in tutte le sue sfumature, ed essere in grado di inserirla in una rete semantica formata dalle associazioni con altre parole.

Il modello di Nation, dunque, inserisce la conoscenza dei rapporti associativi fra le parole (*collocation*)[6] fra gli elementi indispensabili per raggiungere la competenza lessicale (si vedano anche Faerch *et al.*, 1984; Nattinger, DeCarrico, 1992; Lewis, 1993).
In Nation e Webb (2011: 190; Tabella 1), lo schema di Nation (2001) viene adattato alle unità polirematiche, considerate quindi combinazioni lessicali indivisibili, la cui conoscenza, secondo gli autori, si basa su parametri simili a quelli già considerati per le singole parole:

Tabella 1. *Schema proposto da Nation e Webb (2011) sulla conoscenza delle unità polirematiche*

What is involved in knowing a multiword unit (MWU)[7]

Form	Spoken	R	What does the MWU sound like?
		P	How is the MWU pronounced?
	Written	R	What does the MWU look like?
		P	How is the MWU written and spelled?
	Word parts	R	What words are recognizable in this MWU?
		P	What words are needed to express the meaning?
Meaning	Form and meaning	R	What meaning does the MWU signal?
		P	What MWU can be used to express this meaning?
	Concepts and referents	R	What is included in the concept?
		P	What items can the concept refer to?
	Associations	R	What other words or MWUs does this make us think of?
		P	What other words or MWUs could we use instead of this one?
Use	Grammatical functions	R	In what pattern does the MWUs occur?
		P	In what pattern must we use this MWU?

[6] L'idea di associazione contenuta nello schema è ampia e fa riferimento sia alla sfera semantica sia a quella più strettamente grammaticale.
[7] Se si considerano le strutture grammaticali e gli elementi collocati interni alle unità polirematiche, gli autori aggiungono le domande: «What different patterns can occur within this multiword unit? […] What different collocates can occur within this multiword unit?» (Nation, Webb, 2011: 189).

Use	Collocations	R	What words, MWUs, or types of MWUs occur with this one?
		P	What words, MWUs, or types of MWUs must we use with this one?
	Constraints on use (register, frequency, etc.)	R	When, where and how often would we expect to meet this MWU?
		P	When, where and how often can we use this MWU?

Lo schema di Nation e Webb fa riferimento alle unità lessicali formate da più parole. L'inclusione delle collocazioni fra queste non è unanimemente condivisa: le combinazioni fra gli elementi lessicali che compongono le collocazioni, infatti, sono privilegiate, ma non obbligatorie (Simone, 1996: 434). Di fatto, il maggior ostacolo per l'acquisizione delle collocazioni da parte di un parlante non nativo è rappresentato dalla difficoltà nel prevedere restrizioni combinatorie istituzionalizzate dall'uso (Jezek, 2005).

La competenza collocazionale e la competenza fraseologica in generale rappresentano una parte imprescindibile della competenza socio-pragmatica e risultano fondamentali per lo sviluppo di una buona capacità espressiva. Laufer e Waldmann (2011: 649) indicano tre metodologie di studio della conoscenza e dell'uso delle collocazioni negli apprendenti di una L2:

1. l'analisi degli errori applicata a testi previamente selezionati;
2. l'elicitazione attraverso test mirati in cui l'apprendente debba produrre o riconoscere collocazioni specifiche;
3. l'analisi di grandi corpora di apprendenti.

In questo contesto ci soffermeremo sul secondo punto, corrispondente alla metodologia utilizzata nel test oggetto del presente studio.

Laufer e Waldmann (2011), Pérez Serrano (2017) e Gallina (2022) realizzano una panoramica sulle tecniche di elicitazione proposte da diversi autori a partire dagli anni '90 del secolo scorso, classificando i test in base alla componente della competenza collocazionale oggetto di valutazione.

I test proposti da Biskup (1992) e Hasselgren (1994) sono stati disegnati per valutare la capacità produttiva di collocazioni in inglese attraverso la traduzione dalla L1 degli apprendenti (polacco e tedesco nel primo caso e norvegese nel secondo), mentre la ricerca di Bahns e Eldaw (1993) sulla conoscenza delle collocazioni verbo + nome (oggetto) in inglese, realizzata su 58 studenti tedeschi, si basa sulla combinazione di due test, uno di traduzione e un cloze-test in cui è stato eliminato il collocato verbale. Simile è la struttura dei test utilizzati da Farghal e Obiedat (1995, che si proponevano di analizzare le strategie messe in atto dagli apprendenti nell'uso delle collocazioni in L2) e dei test proposti da Gitsaki (1999).

Per quanto riguarda la conoscenza ricettiva, in Gyllstad (2005) si combinano due tipologie di test: negli item del primo (COLLEX), vengono proposte due combinazioni di parole, una pseudo-collocazione e una collocazione che deve essere identificata dagli informanti; nel secondo test, COLLMATCH, si richiede di individuare tutte le combinazioni ammissibili fra una lista di tre verbi e sei nomi. In entrambi i test, le combinazioni sono decontestualizzate. Di seguito si riportano gli esempi citati da Gyllstad (2005):

COLLEX (Gyllstad, 2005: 14)

1)	tell a prayer	say a prayer	
2)	pay a visit	do a visit	
3)	run a diary	keep a diary	
4)	do a mistake	make a mistake	

COLLMATCH (Gyllstad, 2005: 16)

	charges	patience	weight	hints	anchor	blood
drop						
lose						
shed						

Nelle versioni successive (Gyllstad, 2007), il formato del COLLEX viene reso a scelta multipla a tre opzioni, mentre il formato del COLLMATCH diventa un test sì/no in cui il candidato deve indicare se le combinazioni proposte esistono nella lingua inglese:

COLLEX 5 (estratto da Gyllstad, 2007: 306)

				a	b	c
1	a. do damage	b. make damage	c. run damage			
2	a. turn out a fire	b. put out a fire	c. set out a fire			

COLLMATCH 3 (estratto da Gyllstad, 2007: 310)

1 have a say	2 lose sleep	3 do justice	4 draw a breath	5 turn a reason
☐ yes	☐ yes	☐ yes	☐ yes	☐ yes
☐ no	☐ no	☐ no	☐ no	☐ no

Il test CONTRIX disegnato da Revier (2009: 129) si propone di valutare la conoscenza dell'intera collocazione e non solo la capacità di combinare un collocato alla base data. Nel CONTRIX vengono forniti piccoli testi da cui è stata omessa la collocazione che deve essere individuata scegliendo i costituenti fra quelli proposti, che sono stati organizzati in colonne, come in una matrice:

The only way to win a friend's trust is to show that you are able to	tell	a/an	joke
	keep	the	secret
	take		truth

Hargreaves (2000: 220) descrive alcune tipologie di test sulle collocazioni somministrati presso il dipartimento *English as a Foreign Language* dello UCLES (*University of Cambridge Local Examinations Syndicate*). Gli esempi proposti si focalizzano sulla capacità dell'apprendente di mettere in relazione la base con il collocato. Il primo esempio si

avvicina alla tipologia di test usata nel nostro studio. Si tratta infatti di item in cui viene fornito un piccolo testo da cui è stata eliminata una delle due parti che compongono la collocazione e in cui vengono offerte quattro possibilità di scelta. Secondo Hargreaves (2000: 220), questa formula permette all'apprendente non solo di individuare la collocazione appropriata ma anche di escludere consapevolmente le combinazioni sbagliate.

Un'altra possibile modalità per testare la conoscenza delle combinazioni proposta da Hargreaves (2000: 221) consiste nell'offrire una serie di collocazioni in cui occorra la stessa parola, anche se per la validità di questo tipo di test è necessario che tutte le combinazioni presentino lo stesso livello di difficoltà:

> *Circle the word which fits in all three sentences:*
>
> A fashion B opinion C feeling D will
>
> a) You cannot simply come into an existing situation and impose your _____ on everyone like that.
>
> b) Though he may have good reasons for introducing such measures, popular _____ is likely to prevent them from working.
>
> c) She may insist on such a dress code in the office, but whether it's correct to do so is a matter of _____

3. Realizzazione del COLL-IT, un test sulle collocazioni verbo-nome

Uno dei primi passi nella realizzazione di un test consiste nella definizione del costrutto, ovvero nella selezione dei tratti che possono essere considerati rappresentativi della totalità della competenza e che saranno poi sottoposti a valutazione (Barni, 2023). Per realizzare il COLL-IT, la competenza collocazionale è stata operazionalizzata come la capacità di associare a una determinata base (per esempio, *attenzione*) il collocato appropriato (*prestare*). Tra i formati di test più utilizzati a tale scopo vi è il test di completamento (Gallina, 2022). Si tratta di un tipo di test spesso usato per valutare la comprensione del testo o la produzione di strutture target in cui si chiede al candidato di completare un brano, una frase, un sintagma o, appunto, una collocazione con la (o le) parola mancante. In particolare, si è optato per un test di completamento a scelta multipla. Il formato a risposta chiusa è stato preferito rispetto a quello a risposta aperta per diverse ragioni. In primo luogo, consente una correzione più rapida e oggettiva. Nei test a risposta chiusa, infatti, la risposta corretta è univoca, mentre nei test a risposta aperta la valutazione può dipendere dalle scelte del correttore, in particolare per quanto riguarda eventuali errori ortografici o di morfosintassi. In secondo luogo, i test a risposta chiusa riducono la variabilità delle risposte. Nelle domande a scelta multipla, infatti, le opzioni di risposta sono limitate e non c'è il rischio che gli studenti utilizzino parole che, pur essendo accettabili nel contesto, non costituiscono una collocazione (Gallina, 2022). Infine, la somministrazione di un test a risposta chiusa risulta più pratica, soprattutto quando, come nel caso del COLL-IT, il numero di item è elevato. Un test a scelta multipla, che richiede semplicemente di selezionare il verbo corretto, comporta infatti un minore carico cognitivo rispetto a un test che richiede la produzione del verbo mancante.

Nel COLL-IT il candidato deve quindi selezionare il verbo giusto per completare la collocazione tra cinque opzioni di risposta possibili di cui una sola corretta sul piano formale e accettabile nel contesto. Proprio per questo motivo, anziché costruire item decontestualizzati e chiedere al candidato di selezionare il verbo da associare ad una determinata base presentando la collocazione in isolamento, si è deciso di inserire le

collocazioni all'interno di brevi frasi che forniscano un contesto d'uso sufficientemente ampio da evitare ambiguità nell'interpretazione del significato dell'item.

Una volta stabilito il formato del test, sono state selezionate le collocazioni target. Per individuare le collocazioni corrispondenti ad ogni livello (B1, B2 e C[8]), è stato preso come riferimento il sillabo delle collocazioni verbo-nome italiane (La Russa *et al.*, 2023). Il sillabo, realizzato a partire dal corpus di apprendenti CELI (Spina *et al.*, 2022) e dal corpus di nativi PEC (Spina, 2014), include 953 combinazioni verbo-nome (oggetto). Facendo ricorso sia a dati empirici, come la frequenza della collocazione nelle produzioni dei nativi e il suo numero di occorrenze nelle produzioni degli apprendenti, sia a giudizi qualitativi basati sulle liste lessicali dei livelli da A1 a B2 del *Profilo della lingua italiana* (Spinelli, Parizzi, 2010) e l'argomento a cui si riferisce la collocazione, 221 combinazioni sono state assegnate al livello B1, 365 al livello B2 e 358 al livello C (per maggiori dettagli si veda La Russa *et al.*, 2023). Ai fini del test, sono state selezionate 33 collocazioni per ogni livello utilizzando la funzione CASUALE di Excel.

Definiti i target, sono stati redatti gli item. Per ogni collocazione sono state ricercate le concordanze[9], esempi d'uso della struttura target in brevi frasi, all'interno del corpus *Italian corpus for SKELL* (*itSKELL*)[10], appositamente pensato per apprendenti di italiano. La frase da usare come stimolo è stata selezionata usando la funzione *Good dictionary examples* (*GDEX*, Kilgarriff *et al.*, 2008)[11] di *Sketch Engine*[12], che mostra le migliori concordanze per la collocazione ricercata. Questa procedura ha permesso di selezionare frasi semplici per fare in modo che la comprensione dello stimolo non risultasse difficile per gli informanti e che potessero invece concentrarsi sul task di completamento della collocazione. Per facilitare la comprensione degli apprendenti di livello più basso (B1) sono state privilegiate frasi formate da vocaboli e strutture linguistiche adatti per quel livello e con il verbo della collocazione al modo infinito o indicativo.

Da ogni frase-stimolo è stato poi rimosso il verbo della collocazione target e sono state date cinque opzioni di risposta possibili di cui una risposta esatta, tre distrattori e l'opzione "non lo so" da selezionare nel caso in cui non si conosca la risposta corretta per evitare risposte casuali. Per i distrattori sono stati inizialmente selezionati:

- un verbo foneticamente simile alla risposta corretta. Per esempio, *celare* è stato usato come verbo foneticamente simile a *cercare*;
- un verbo semanticamente simile. Per esempio, *provare* è stato usato come sinonimo di *cercare*;
- un verbo supporto tra *dare*, *fare*, *prendere*, *mettere* e *avere*.

[8] Inizialmente era presente una distinzione tra i livelli C1 e C2, tuttavia, ci si è resi conto che tale distinzione risultava spesso arbitraria e difficile da determinare in modo oggettivo. Per questa ragione è stato creato un unico livello C. Per maggiori informazioni si veda La Russa *et al.* (2023).

[9] Per effettuare la ricerca delle concordanze delle collocazioni verbo-nome è stata utilizzata una query CQL (*Corpus Query Language*). CQL è un linguaggio che consente di effettuare ricerche avanzate all'interno di corpora linguistici per trovare specifici pattern grammaticali o lessicali (per esempio, le collocazioni verbo-nome). La query a cui si è fatto ricorso è la seguente: [lemma = "verbo target"] []? [lemma = "nome target"], per esempio: [lemma = "prendere"] []? [lemma = "decisione"].

[10] SKELL *(Sketch Engine for Language Learning)* è un'interfaccia di interrogazione, selezione e visualizzazione che, grazie a un algoritmo, seleziona 40 esempi adatti per apprendenti in termini di leggibilità, ovvero privi di subordinate troppo lunghe, lessico tecnico-specialistico, riferimenti anaforici o altri deittici difficili da comprendere, ecc. In questo caso l'interrogazione viene effettuata all'interno del corpus *itTenTen 2016,* composto da testi in italiano raccolti da Internet. Per maggiori informazioni si veda https://www.sketchengine.eu/itskell-italian-corpus/#toggle-id-2.

[11] *GDEX* assegna un punteggio a ogni frase sulla base della sua lunghezza (lunghezza minima e massima) e della frequenza delle parole che la compongono. Le frasi con il punteggio più alto sono visualizzate come primi risultati di una concordanza.

[12] https://www.sketchengine.eu/.

Prima della somministrazione ufficiale, tre pre-tester che avevano l'italiano come L2 hanno risposto al questionario per verificare che non ci fossero anomalie o errori. Uno dei pre-tester ha riferito di esser riuscito a trovare più facilmente la risposta corretta per via delle due parole foneticamente simili tra le opzioni. Per questo motivo, il distrattore foneticamente simile alla risposta corretta è stato sostituito da un verbo casuale (per esempio, *celare*, foneticamente simile alla risposta esatta *cercare*, è stato sostituito da *aprire*).

Sono stati scelti come distrattori unicamente verbi transitivi per fare in modo che ogni opzione di risposta mantenesse lo stesso pattern della risposta corretta, ovvero verbo-complemento oggetto.

Il seguente è un esempio di item presente nel test:

> Esempio (1):
> *Scegli il verbo corretto per completare la frase. Se non conosci la risposta, scegli 'non lo so'.*
> Una giuria internazionale _____ il premio al miglior film.
> - noterà (0 punti)
> - prescriverà (0 punti)
> - metterà (0 punti)
> - assegnerà (1 punto)
> - non lo so (0 punti)

Gli item sono stati inseriti in un modulo Google. Le opzioni di risposta sono state ordinate in modo casuale per ogni partecipante al test ed è stato attribuito un punto per ogni risposta corretta e zero punti per le risposte errate e per l'opzione "non lo so". La risposta è stata resa obbligatoria per ogni domanda.

Dato che il COLL-IT è costituito da molti item (99) e che, quindi, la stanchezza e l'ordine di presentazione degli item avrebbero potuto incidere sulla correttezza delle risposte fornite dagli informanti, gli item non sono stati ordinati per livello (per esempio prima quelli corrispondenti al livello B1, poi quelli del B2 e infine quelli di livello C) ma è stata usata la funzione CASUALE di *Excel* per collocarli all'interno del test in ordine casuale ma uguale per ogni candidato.

Prima del test sulle collocazioni è stata inserita una sezione contenente 12 domande volte a raccogliere informazioni sul profilo degli informanti: età; L1; altre lingue nel repertorio verbale; percorso di studio/apprendimento dell'italiano, ovvero luogo in cui è stato studiato, livello dell'ultimo corso frequentato, livello percepito, durata di eventuali soggiorni in Italia e utilizzo dell'italiano al di fuori del corso di lingua.

Allo scopo di misurare in maniera quanto più affidabile la competenza collocazionale di apprendenti di italiano L2/LS di livello intermedio/avanzato (B1, B2 e C) è stato così creato il COLL-IT, un test di completamento a risposta multipla composto da 33 item per ognuno dei livelli coinvolti, per un totale di 99[13].

4. Somministrazione del test

La somministrazione del COLL-IT ha coinvolto 103 candidati tra fine maggio e inizio giugno 2023, di cui 91 studenti dei corsi di laurea in *Lenguas Modernas* e *Traducción e Interpretación* dell'Università di Granada (UGR, Spagna) e 12 studenti del comitato di Granada della Società Dante Alighieri. Fra i candidati iscritti alla UGR, 32 studiano italiano come prima lingua straniera e 59 come seconda lingua straniera.

[13] L'intero test è consultabile al link:
https://osf.io/f7ujw/?view_only=47f3c1c3c024476d8a406bb23f89ba99.

Il livello di uscita previsto dai corsi di provenienza degli studenti della UGR corrisponde a B1 (*Italiano Lengua C2*), B1.2 (*Italiano Intermedio 2*, *maior e minor*), B2 (*Italiano Avanzado 2*) e B2.2 (*Italiano Lengua C6* e *Italiano Superior 2*). Il livello degli studenti dei corsi di italiano della Società Dante Alighieri va dal B2 al C2. Dato che la somministrazione del test è avvenuta durante l'ultima settimana del secondo semestre, è ragionevole considerare che il livello dei candidati e quello previsto alla fine del programma del corso coincidano.

In totale, il test è stato realizzato da 60 studenti di livello B1, 29 di livello B2 e 14 di livello C. Il COLL-IT è stato somministrato in aula durante l'orario dei corsi ai gruppi di *Italiano Lengua C2*, *Italiano Intermedio 2 lengua maior e mino*r, *Italiano Avanzado 2* e *Italiano Lengua C6*, mentre gli studenti di *Italiano Superior 6* e quelli della Società Dante Alighieri, per questioni di incompatibilità oraria, lo hanno realizzato al di fuori dell'orario di lezione, in momenti diversi.

5. Domande di ricerca e metodologia

Il presente studio si propone di verificare alcune ipotesi che verranno esplicitate in questo paragrafo. L'obiettivo principale risiede nella ricerca e interpretazione di evidenze di affidabilità e di validità del COLL-IT e del costrutto di competenza collocazionale, nonché nella verifica del funzionamento degli item che compongono il test. Le domande di ricerca che hanno guidato le analisi sono:

1. In che misura i punteggi ottenuti dai partecipanti alla somministrazione del COLL-IT possono essere ritenuti affidabili?
2. Stando ai punteggi ottenuti dai candidati, quali evidenze vi sono sulla capacità del COLL-IT di discriminare i diversi livelli di competenza linguistica dei candidati?
3. Alla luce della strutturazione interna del test, quali evidenze di validità del costrutto di competenza collocazionale forniscono i punteggi ottenuti nella somministrazione?
4. Quali informazioni si ottengono effettuando un'analisi delle prestazioni dei singoli item, in termini di difficoltà dei quesiti e capacità di discriminazione dei candidati?

Le analisi dei punteggi e degli item del COLL-IT sono state condotte utilizzando i metodi proposti dalla teoria classica del testing o *Classical Test Theory* (CTT), che è stata definita come «[u]na teoria di misurazione che consiste in una serie di supposizioni sulle relazioni fra i punteggi reali e i fattori che possono averli influenzati, chiamati comunemente errori» (ALTE Members, 1998: 314). Difatti, uno degli intenti principali della CTT è quello di ridurre il più possibile gli errori di misurazione, massimizzando dunque l'affidabilità dei punteggi, ossia «[l]a stabilità delle misure ottenute da un test» (ALTE Members, 1998: 284), che è una delle evidenze e qualità necessarie che concorrono alla validità di un test (Weir, 2005). Nel modello di analisi proposto dalla CTT, uno degli assunti fondamentali sostiene che il punteggio osservato (*observed score*) in un test sia composto da due elementi indipendenti: un punteggio reale (*true score*) e un punteggio erroneo (*error score*) (cfr. Bachman, 1990). Da una parte, il punteggio reale corrisponde effettivamente al punteggio da attribuire all'abilità latente che si intende misurare, ovvero al «punteggio che il candidato otterrebbe se non vi fossero stati errori al momento della verifica o dell'attribuzione del punteggio» (ALTE Members, 1998: 309). Dall'altra, il punteggio erroneo consiste in errori di misurazione non sistematici derivanti da fattori estranei all'abilità che si intende misurare, che per assunto sono ritenuti casuali (*random*). Analogamente, la varianza[14] di un insieme di punteggi osservati (*observed score variance*) – per

[14] Per varianza si intende la «[m]isura della dispersione di una serie di punteggi. Più grande è la varianza, più lontani risultano i punteggi dalla media» (ALTE Members, 1998: 320).

esempio l'insieme dei punteggi ottenuti da un gruppo di candidati che partecipa alla somministrazione di un test – è costituita dalla componente della varianza dei punteggi reali (*true score variance*) e dalla componente della varianza degli errori casuali del gruppo (*random error variance*).

Per rispondere alla domanda di ricerca 1, che mira a indagare l'affidabilità del test COLL-IT, verrà impiegato un metodo basato sulle varianze dei punteggi dei singoli item, trattati come misure parallele e indipendenti, e la varianza dei punteggi totali ottenuti dai candidati (Bachman, 2004). Attraverso il coefficiente di affidabilità *a*, o *Alfa di Cronbach* (Cronbach, 1951), verrà calcolata la stima di omogeneità degli item, la cosiddetta *consistenza interna* (*internal consistency*) dello strumento di verifica, ossia «[q]uella caratteristica di un test, rappresentata dal grado di corrispondenza fra i punteggi ottenuti dai candidati nei singoli item del test e il punteggio complessivo» (ALTE Members, 1998: 291). Per semplificare l'interpretazione dei valori stimati da *a*, che teoricamente possono andare da -1 a +1, DeVellis (1991) e DeVellis, Thorpe (2022: 130) propongono dei descrittori per diversi *range* di stima: i valori di *a* minori di 0.60 sono da considerarsi «*unacceptable*», da 0.60 a 0.65 «*undesirable*», 0.65-0.70 «*minimally acceptable*», 0.70-0.80 «*respectable*», 0.80-0.90 «*very good*», mentre quelli oltre 0.90 sono da considerarsi eccellenti (DeVellis, 1991: 85) o addirittura candidati per la rimozione degli item non sufficientemente performanti ai fini di dell'ottimizzazione di uno strumento per successive somministrazioni (DeVellis, Thorpe, 2022: 131-132). L'ipotesi sull'affidabilità del COLL-IT è che il test presenti un valore *a* elevato, dato che questo dipende, oltre che dal buon funzionamento degli item, anche dalla quantità di quesiti che compongono il test stesso (Bachman, 1990): in questo caso, 99 item possono essere considerati un numero consistente di quesiti.

L'impiego del coefficiente *a* come unico indicatore di affidabilità, tuttavia, è stato criticato a più riprese (si veda per es. Sijtsma, 2009; Dunn *et al.*, 2014). Infatti, il coefficiente *a* si basa su assunti la cui violazione potrebbe causare alterazioni nei valori di stima (Kelley, Cheng, 2012); inoltre, secondo i detrattori di questo metodo, l'interpretazione di tali valori è stata spesso operata in maniera eccessivamente semplificata (DeVellis, Thorpe, 2022). Al fine di arginare alcune delle limitazioni di *a*, è stata proposta una tecnica di ricampionamento casuale con rimpiazzo, nota con il nome di *bootstrapping* (Dunn *et al.*, 2014). Il metodo del *bootstrapping* compie il calcolo del coefficiente *a* su molteplici ricampionamenti casuali, generando un campione di stime. Replicando questo procedimento migliaia di volte è possibile calcolare e riportare degli intervalli di confidenza del 95% della stima di *a*, rendendo più solida sia la stima stessa sia le interpretazioni che ne conseguono circa l'affidabilità del test.

Le domande di ricerca 2 e 3 si propongono di indagare aspetti concernenti la validità del test, che può essere definita come «[l]a misura in cui i punteggi di un test rendono possibili delle inferenze appropriate, significative ed utili, in base allo scopo del test stesso» (ALTE Members, 1998: 318). Un numero cospicuo di studiosi oggi tende a definire la validità come una qualità del test composta da numerosi aspetti diversi, i quali concorrono in maniera complementare e non alternativa alla fondatezza delle interpretazioni dei punteggi. Tuttavia, non potendo approfondire in questa sede il concetto di validità[15], è possibile fare brevemente riferimento a Weir (2005), in cui si asserisce che la validità è una forma di valutazione in cui vengono impiegati metodi sia qualitativi sia quantitativi per generare delle evidenze (a priori e a posteriori rispetto a una somministrazione) a sostegno del costrutto operazionalizzato e in supporto delle inferenze operate sui punteggi del test (Weir, 2005: 11-40).

[15] Per dettagli, si veda per es. Messick (1989), Weir (2005), Bachman, Palmer (2010), Kane (2012), Chapelle (2012; 2021).

Per quanto concerne la metodologia di analisi relativa alla domanda di ricerca 2, si è scelto di impiegare il metodo di confronto tra gruppi non equivalenti, composti a priori sulla base di un criterio esterno (cfr. Henning, 1987: 98; Bachman, 2004: 290), il livello di competenza linguistico-comunicativa dell'apprendente. La capacità di un test di discriminare tra le performance di candidati con diversi livelli di abilità, infatti, è strettamente legata alla validità dei punteggi di un test (cfr. Gyllstad, 2009) e, di conseguenza, delle interpretazioni basate su di essi. Nello specifico, si cercherà di reperire evidenze circa la capacità del test di competenza collocazionale di riflettere attraverso i suoi punteggi anche il livello di competenza linguistica QCER dei candidati. L'ipotesi è che vi siano evidenze di uno sviluppo parallelo tra la competenza collocazionale così come operazionalizzata in questo studio e la competenza linguistico-comunicativa dei candidati coinvolti nella somministrazione e, perciò, che all'aumentare medio del punteggio dei candidati del COLL-IT corrisponda tendenzialmente un più alto livello di competenza linguistica generale.

Per quanto attiene alla domanda di ricerca 3, si cercherà di reperire evidenze di validità del costrutto di competenza collocazionale, ovvero di prove a supporto dell'ipotesi secondo la quale degli item basati su un sillabo di collocazioni di livello B1, B2 e C genererebbero punteggi mediamente inferiori man mano che vengono testate collocazioni target di livello QCER superiore. In questo caso, l'ipotesi è che vi sia un incremento progressivo della competenza collocazionale di apprendenti di italiano man mano che affrontano item che presuppongono una maggiore difficoltà di apprendimento delle collocazioni target. Perciò l'aspettativa è che i candidati ottengano punteggi mediamente più alti nella sezione di item di livello QCER più basso (B1), e che, all'aumentare della difficoltà ipotizzata degli item, quindi nella sezione B2 del test e ancor più in quella C, i candidati ottengano punteggi medi più bassi. In particolare, si ipotizza che ciò possa avvenire soprattutto all'interno dei singoli gruppi di competenza dei candidati (gruppo B1, B2 e C).

La domanda di ricerca 4 si pone come obiettivo quello di indagare il funzionamento e le prestazioni degli item che compongono il COLL-IT. Ciò verrà eseguito attraverso un metodo chiamato *analisi degli item*, che prevede «[u]na descrizione della prestazione fornita dai candidati nei singoli item del test, [...] abitualmente fatta ricorrendo ad indici [...] quali l'indice di facilità o l'indice di discriminazione» (ALTE Members, 1998: 285). Per indice di facilità (*item facility index*, d'ora in poi IF) si intende «[l]a proporzione di risposte corrette ad un item, trascritta in una scala da 0 a 1 [...] [oppure] in percentuale» (ALTE Members, 1998: 301). L'importanza delle informazioni derivanti dell'IF risiede nel fatto che si tratta di un indicatore che consente di comprendere se un determinato item è adatto per il gruppo a cui viene somministrato il test (McNamara, 2000: 60). Infatti, pur in mancanza di una regola generale sulle soglie minime e massime di accettabilità di IF, intimamente legate allo scopo, al gruppo target e all'uso che si intende fare del test (Green, 2013: 27), gli studiosi sono concordi sul fatto che il valore ideale di IF di un item deve tendere a 0.50. In altre parole, un item ben costruito tende a produrre una risposta corretta da parte del 50% del gruppo target. In questo studio, si è deciso di considerare accettabile un intervallo di valori di IF che va da 0.20 a 0.80 (Green, 2013: 26-27). Per indice di discriminazione dell'item (*item discrimination index*, d'ora in avanti ID) si intende «[i]l potere che ha un item di distinguere i candidati più bravi dai candidati meno bravi» (ALTE Members, 1998: 295). In sostanza, se un determinato item possiede un buon indice di discriminazione, i candidati che hanno ottenuto un punteggio elevato sull'intero test dovrebbero in media fornire risposte corrette al quesito rispetto a quanto non facciano mediamente i candidati meno performanti. Perciò, se un item è difficile (cioè riporta un IF basso) ci si aspetta che il gruppo di candidati più bravi risponda meglio del gruppo di candidati meno bravi; se ciò non avviene, il valore ID dell'item si abbassa, rispecchiando

l'incapacità dell'item stesso di discriminare in modo affidabile un candidato più bravo da uno meno bravo. Analogamente, se un item è facile (valore IF alto) è probabile che riporti un valore ID basso, dal momento che probabilmente non è in grado di discriminare un candidato bravo da uno meno bravo (entrambi danno risposte simili). Sulla base dei valori ID degli item, si possono individuare gli item più indicativi nel riflettere l'abilità che si intende misurare e si possono circoscrivere quelli che richiedono interventi di revisione o eventuale eliminazione (McNamara, 2000: 61). In questo studio si è scelto di utilizzare un coefficiente di discriminazione molto diffuso, l'indice di correlazione punto-biseriale (*point biserial correlation*), ovvero il calcolo dell'indice di correlazione tra il punteggio totale del test e le risposte agli item (Henning, 1987: 52), nella sua versione corretta (*corrected item-total correlation*; Green, 2013). Per l'interpretazione di ID, i cui valori vanno da -1 a +1, Ebel (1979) e poi Popham (2000) propongono il seguente schema (Tabella 2):

Tabella 2. *Proposta di interpretazione dei livelli di discriminazione da parte di Popham (2000), riportata in Green (2013: 29)*

.40 and above	Very good items.
.30 to .39	Reasonably good items but possibly subject to improvement.
.30 to .39	Reasonably good items but possibly subject to improvement.
.19 and below	Poor items, to be rejected or improved by revision.

Altri studiosi ritengono accettabili i valori ID superiori a 0.25[16] (Henning, 1987; Green, 2013) e in questo studio ci si atterrà a tale soglia minima, pur facendo riferimento allo schema interpretativo di Ebel (1979) e Popham (2000). Al fine di verificare più a fondo i livelli di discriminazione e prima di procedere alla eventuale revisione o eliminazione di item poco performanti, si è scelto di condurre anche la cosiddetta *analisi dei distrattori* (ovverosia delle opzioni non corrette) degli item che presentino ID inferiori alla soglia minima. Il metodo scelto si basa sulla suddivisione in terzili della distribuzione dei punteggi totali ottenuti dai partecipanti al test (Allen, Yen, 2002: 123). Perciò, sono state individuate tre sezioni della popolazione di candidati, composte ognuna da circa il 33% di essi: la sezione del gruppo di candidati con punteggi totali bassi (SPB) che conta 35 candidati, la sezione con punteggi medi (SPM) che ne conta altri 35, e quella dei punteggi alti (SPA) che assomma 33 partecipanti, per un totale di 103. Per ogni opzione di risposta di un determinato item sotto indagine si prenderanno in considerazione la *frequenza* e la *proporzione* in percentuale con cui è stata selezionata dai componenti di ciascuna sezione di punteggi (SPB, SPM, SPA), al fine di fornire un'interpretazione plausibile circa il funzionamento deficitario dell'item stesso.

6. ANALISI DEI DATI E DISCUSSIONE DEI RISULTATI

In questo paragrafo verranno presentate le analisi dei punteggi derivanti dalla somministrazione del COLL-IT e ne verranno discussi i risultati. Le analisi quantitative sono state condotte per mezzo del software di analisi statistica *RStudio* (versione 1.4.1106; RStudio Team, 2021), dei pacchetti *boot* (Canty, Ripley, 2022), *car* (Fox, Weisberg, 2019), *CTT* (Willse, 2018), *dunn.test* (Dinno, 2017), *ez* (Lawrence, 2016) e *psych* (Revelle, 2017), mentre i grafici sono stati elaborati utilizzando Microsoft Excel.

[16] Per i test *high stakes* la soglia minima per i valori ID degli item è comunemente fissata a 0.30, per quelli *low stakes* 0.25 (Green, 2013).

Le statistiche descrittive dei punteggi dei candidati sull'intero test, suddivisi per le sezioni del test e per i livelli dei candidati, sono riassunte nella Tabella 3. Il punteggio medio ottenuto dai 103 candidati è di 70,38 su 99 punti possibili, con deviazione standard (d'ora in poi DS) di 16,69, mentre il punteggio minimo totalizzato è stato 26/99 e quello massimo di 99/99. Nella sezione B1 del test, i candidati hanno totalizzato in media 24,95 punti su 33 (DS 5,57), nella sezione B2 23,52/33 (DS 6,04) e in quella C 21,9/33 (DS 5,94).

Se si prendono in considerazione singolarmente i gruppi di candidati di livello diverso, si nota che i punteggi complessivi aumentano sensibilmente all'aumentare del livello dei candidati (B1: 61,6/99, DS 15,07; B2: 77,97/99, DS 7,78; C: 92,29/99, DS 4,92). Inoltre, anche le prestazioni sulle sezioni diverse del test da parte dei candidati di uno stesso livello indicano che i punteggi tendono in media a diminuire all'aumentare del livello QCER della sezione del test (punteggio medio sezione B1 > sezione B2 > sezione C). In § 6.2, si verificherà se queste differenze sono statisticamente significative o meno.

Tabella 3. *Punteggi medi e deviazione standard (DS) sull'intero test e suddivisi per sezioni del test e per livello dei candidati*

Sezione test (numero di item)	**Candidati B1 (n = 60)**	**Candidati B2 (n = 29)**	**Candidati C (n = 14)**	**Totale candidati (n = 103)**
	Media punteggi (DS)			
Sezione B1 (k=33)	22,07 (5,2)	27,72 (2,75)	31,57 (1,09)	24,95 (5,57)
Sezione B2 (k=33)	20,47 (5,48)	26,17 (3,43)	31,14 (1,79)	23,52 (6,04)
Sezione C (k=33)	19,07 (5,5)	24,07 (3,1)	29,57 (2,77)	21,9 (5,94)
Intero test (k=99)	**61,6 (15,07)**	**77,97 (7,78)**	**92,29 (4,92)**	**70,38 (16,69)**

6.1. *Affidabilità del test*

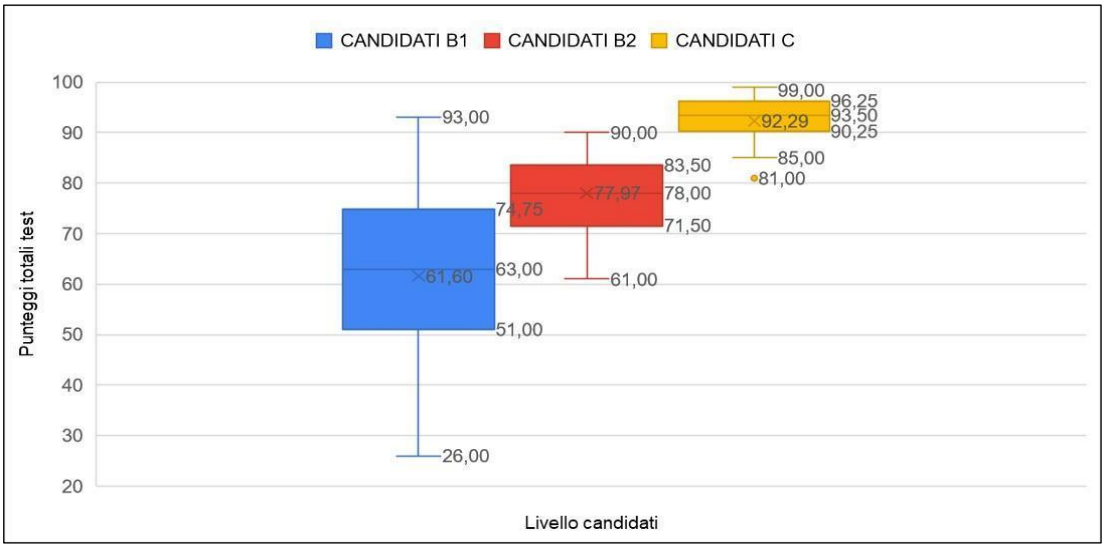
La domanda di ricerca 1 si propone di indagare l'affidabilità del test COLL-IT e il metodo prescelto è quello del computo dell'indice di consistenza interna dei punteggi, l'*Alfa* (*a*) *di Cronbach* (Cronbach, 1951). L'analisi condotta sui punteggi ottenuti dai 103 partecipanti alla somministrazione ha restituito una stima del valore α pari a 0.95, il che può considerarsi un esito auspicabile (Green, 2013), o *excellent*, secondo i descrittori individuati da DeVellis (1991: 85). Inoltre, sono stati calcolati gli intervalli di confidenza della stima attraverso il metodo del *bootstrapping* con 10.000 repliche (Dunn *et al.*, 2014): 0.93 e 0.96. Se si prendono in considerazione i soli quesiti che afferiscono al gruppo di 33 item selezionato dal livello B1 del sillabo (cfr. § 3), la stima di consistenza interna α raggiunge il valore di 0.86, per il gruppo di livello B2 di 0.87, e per il gruppo di fascia C di 0.87.

I risultati riferiti all'intero test sono in linea con le aspettative. Inoltre, anche i valori stimati per ognuna delle sottosezioni del test, che avrebbero potuto rivelarsi carenti perché composti da un numero di item di gran lunga inferiore all'intero test, possono a loro volta essere ritenuti soddisfacenti, o *very good* (DeVellis, 1991: 85; DeVellis, Thorpe, 2022: 131-132). Si può dunque affermare che la somministrazione ha prodotto risultati che sono indice di un ottimo livello di affidabilità dei punteggi, sia per l'intero test sia per le sottosezioni comprendenti item con collocazioni target di livelli diversi.

6.2. *Evidenze di validità del test*

Con il fine di ottenere prime evidenze di validità del test, si è scelto di seguire il metodo di confronto tra gruppi (Henning, 1987: 98). In particolare, la domanda di ricerca 2 è volta a sondare la capacità del COLL-IT di discriminare tra i livelli di competenza linguistica dei candidati (B1, B2, C). Per fare ciò, è stato necessario confrontare le medie dei punteggi totali ottenuti dai candidati dei tre diversi livelli di competenza QCER (media B1: 61,6, DS: 15,07; B2: 77,8, DS: 7,78; C: 92,3, DS: 4,92), attraverso un'analisi di varianza (ANOVA) a una via tra i tre gruppi. Una delle assunzioni dell'ANOVA, oltre alla normalità della distribuzione dei dati, consiste nel fatto che le varianze dei valori dei gruppi siano omogenee. Perciò, per verificare l'omogeneità delle varianze, è stato effettuato il test di Levene sui gruppi di punteggi. Una volta stabilito che le varianze dei tre gruppi non erano equivalenti ($F = 14.17$, $p < 0.001$, $\eta^2 = 0.22$), si è optato per un metodo con caratteristiche simili al test di varianza ANOVA, ovvero il test non parametrico Kruskall-Wallis, che si basa sul confronto delle mediane e dei ranghi (e non delle medie, come nel caso dell'ANOVA) dei gruppi di punteggi in esame (per una rappresentazione, si veda il Grafico 1). I risultati del test Kruskall-Wallis hanno rilevato differenze statisticamente significative tra i punteggi dei tre gruppi di candidati ($\chi^2(2) = 51.62$, $p < 0.001$), con una dimensione dell'effetto molto ampia ($\eta^2 = 0.5$). Per determinare con esattezza tra quali gruppi erano riscontrabili differenze attraverso dei confronti multipli, è stato effettuato il test post-hoc di Dunn che, utilizzando il metodo correttivo di Bonferroni con α pari a 0.017, ha indicato che le differenze di punteggio sono statisticamente significative tra tutti i gruppi (B1-B2: $p < 0.001$; B1-C: $p < 0.001$; B2-C: $p < 0.01$). I risultati delle analisi statistiche hanno quindi confermato che vi sono differenze nei punteggi ottenuti dai candidati di livello diverso e che, come atteso, all'aumentare della competenza linguistica generale dei candidati aumentano in maniera statisticamente significativa anche i punteggi totali ottenuti sul test di competenza collocazionale COLL-IT.

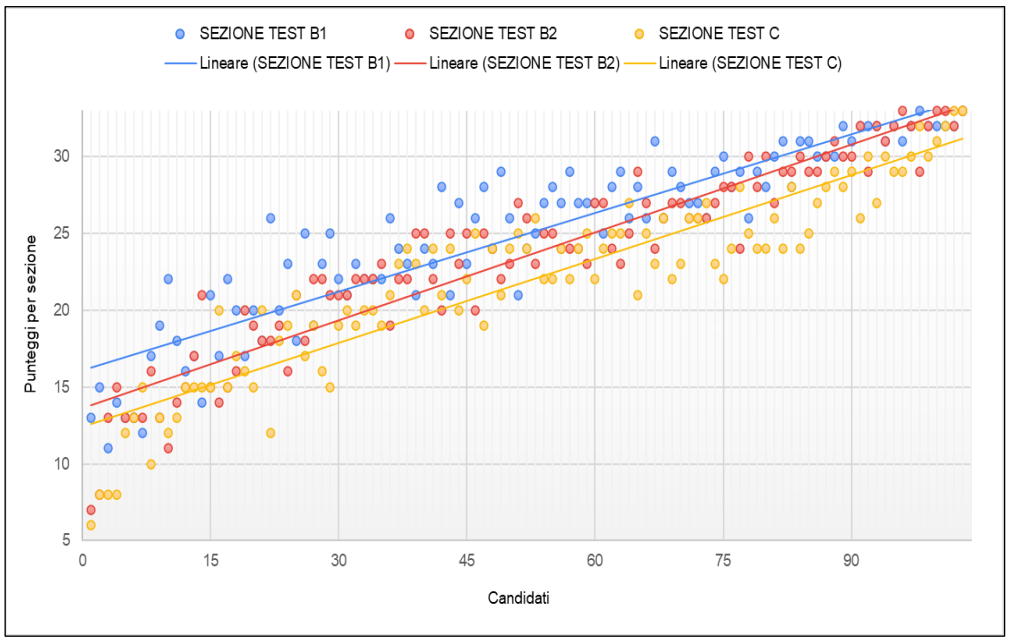
Grafico 1. *Diagramma a scatola e baffi dei punteggi totali dei candidati, raggruppati per livello di competenza QCER*



La domanda di ricerca 3 riguarda le evidenze di validità del costrutto di competenza collocazionale. Come già descritto, il COLL-IT è strutturato in tre sezioni, ognuna composta da 33 item che hanno come target le 33 collocazioni selezionate nei livelli B1,

B2 e C del sillabo cui si è fatto menzione in § 3, per un totale di 99 item. Per tale motivo, l'obiettivo è quello di verificare se i punteggi medi della somministrazione (che, come descritto in § 6.1, diminuiscono all'aumentare del livello QCER delle sezioni del test) differiscono anche dal punto di vista della significatività statistica, per le tre sezioni di item del COLL-IT. Dunque, se si prendono in considerazione le diverse sezioni del test e si analizzano dal punto di vista dei punteggi dell'intera popolazione di candidati (n = 103), si dovrebbero ottenere delle prime evidenze sulla bontà (o meno) del costrutto di competenza collocazionale, così concepito in questo studio, ovvero come capacità di associare a una determinata base il collocato appropriato. A questo scopo, è stato condotto un test di varianza entro casi tra i gruppi dei punteggi di tutti i candidati in ognuna delle sezioni del COLL-IT (media punteggi della sezione del test B1 = 24,95 DS 5,57; B2 = 23,52 DS 6,04; C = 21,9 DS 5,94). Considerando dei punteggi in relazione tra loro, il test di varianza entro casi, oltre all'assunzione di normalità delle distribuzioni, prevede un assunto di sfericità dei dati, ovvero che le covarianze dei gruppi di punteggi siano omogenee. Attraverso un test Shapiro-Wilk ($\alpha = 0.05$) è stata verificata l'assunzione di normalità della distribuzione dei dati ($p = 0.552$), mentre l'assunto di sfericità è stato verificato attraverso il test di Mauchly ($\chi^2(2) = 4.086$, $p = 0.130$, Greenhouse-Geisser $\varepsilon = 0.96$). Il test di Shapiro-Wilk e di Mauchly non hanno rilevato violazioni delle assunzioni di normalità e di sfericità e si è dunque proceduto con il test di varianza entro casi, che ha restituito risultati statisticamente significativi ($F(2, 204) = 48.62$, $p < 0.001$), nonostante una ridotta dimensione dell'effetto ($\eta^2 = 0.04$). Una serie di t-test post-hoc di confronto multiplo tra i gruppi, con correzione di Bonferroni ($\alpha = 0.017$), ha confermato differenze significative tra tutte le coppie di gruppi ($p < 0.001$). Questi risultati confermano le aspettative, dimostrando che i candidati mediamente ottengono punteggi via via inferiori all'aumentare della difficoltà ipotizzata degli item. Più nel concreto, i partecipanti al test hanno dimostrato maggiori difficoltà nel riconoscere correttamente il collocato di una collocazione target estratta da una sezione del sillabo di livello QCER superiore. Le prestazioni dei candidati (riflesse nei punteggi) sono caratterizzate da differenze statisticamente significative e tali risultati costituiscono una prima evidenza preliminare della validità del costrutto di competenza collocazionale.

Grafico 2. *Punteggi parziali (da 0 a 33) ottenuti da ognuno dei 103 candidati in ognuna delle sezioni (B1, B2, C) del test*



Come si può notare nel Grafico 2, che ritrae i punteggi di ogni candidato su ciascuna delle sezioni del COLL-IT, i candidati meno performanti (parte sinistra del grafico) ottengono mediamente un punteggio nettamente superiore nella sezione B1 rispetto alle altre sezioni. Le differenze dei punteggi sembrano assottigliarsi per i candidati mediamente performanti (parte centrale del grafico), fino ad appiattirsi per i candidati più bravi (parte destra del grafico).

Se da una parte l'analisi proposta può essere in qualche modo influenzata dal numero di apprendenti di diversi livelli coinvolti nella somministrazione, queste evidenze paiono indicare che il COLL-IT possa essere adeguato per un impiego con apprendenti di diversi livelli. Inoltre, è possibile raffinare ulteriormente la ricerca di evidenze di validità del costrutto operando confronti tra i punteggi internamente ai singoli livelli. Da un punto di vista acquisizionale, infatti, si ipotizza che i punteggi del COLL-IT forniscano evidenze di validità anche all'interno dei gruppi di punteggi dei candidati di ciascun livello. Per effettuare questa verifica, occorre svolgere dei test di varianza entro casi sui punteggi ottenuti da ciascun gruppo di candidati di livello di competenza differente nelle diverse sezioni del COLL-IT, le quali, per l'appunto, corrispondono a serie di collocazioni che si ipotizza che vengano apprese man mano che si approfondisce la competenza linguistica in italiano come lingua non materna, cioè a livelli QCER differenti, via via superiori (si veda il Grafico 3 per una rappresentazione dei dati).

Per quanto concerne i candidati di livello B1 ($n = 60$), sono stati ottenuti risultati che evidenziano un decremento nei punteggi man mano che la competenza richiesta aumenta, con differenze statisticamente significative ($F(1.79, 105.79) = 23.11$, $p < 0.001$, $\eta^2 = 0.05$)[17] per ognuna delle coppie (B1-B2: $p = 0.001$; B1-C: $p < 0.001$; B2-C: $p < 0.001$). Per quanto riguarda i candidati di livello B2 ($n = 29$), i risultati dello stesso procedimento rilevano un decremento nei punteggi analogo a quanto avvenuto per i candidati del livello precedente ($F(2, 56) = 22.2$, $p < 0.001$ $\eta^2 = 0.19$)[18], ma con una significatività leggermente attenuata nelle differenze tra i gruppi (B1-B2: $p < 0.01$; B1-C: $p < 0.001$; B2-C: $p = 0.001$). Quanto al gruppo di candidati di livello C ($n = 14$), pur essendoci riscontri di un lieve decremento nel punteggio medio nelle tre sezioni all'aumentare della competenza richiesta, con differenze anche significative ($F(2, 26) = 7.82$, $p = 0.002$, $\eta^2 = 0.17$)[19], queste ultime vengono individuate solamente tra i gruppi di item B1 e C ($p < 0.01$), mentre non tra le altre coppie (B1-B2: $p = 0.34$; B2-C: $p = 0.02$)[20].

In linea di massima, le ipotesi riguardanti la domanda di ricerca 3 sono state verificate. Come mostrato, per tutti i livelli dei candidati i punteggi medi sono decresciuti all'aumentare della competenza linguistica QCER assegnata alle collocazioni target delle diverse sezioni di item del test. Ciò è particolarmente evidente per i candidati di livello B1 e di livello B2. Infatti, questi presentano un andamento lineare nei punteggi. Degna di nota è la (lievemente) attenuata significatività nella differenza di punteggio tra la sezione B1 e B2 del COLL-IT per i candidati di livello B2: questo fatto è un probabile indice dell'incipiente livellamento della competenza dei candidati nelle sezioni di item B1 e B2, le cui collocazioni target si suppone vengano apprese con successo una volta raggiunto un pieno livello di competenza B2. Nonostante il numero esiguo di componenti del gruppo di livello C, che impone una valutazione dei risultati ottenuti estremamente prudente, per tale raggruppamento viene confermato il livellamento delle competenze

[17] Assunzione di normalità verificata con il test Shapiro-Wilk ($\alpha = 0.05$), con $p = 0.7379$; assunzione di sfericità verificata con il test di Mauchly $\chi^2(2) = 7.11$, $p = 0.029$, Greenhouse-Geisser $\varepsilon = 0.897$).

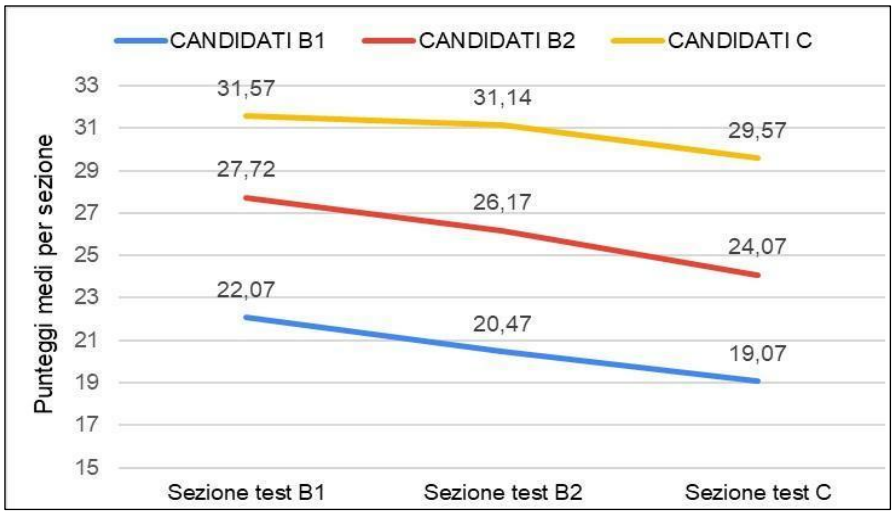
[18] Assunzione di normalità verificata con il test Shapiro-Wilk ($\alpha = 0.05$), con $p = 0.3335$; assunzione di sfericità verificata con il test di Mauchly $\chi^2(2) = 0.443$, $p = 0.801$, Greenhouse-Geisser $\varepsilon = 0.984$).

[19] Assunzione di normalità verificata con il test Shapiro-Wilk ($\alpha = 0.05$), con $p = 0.1928$; assunzione di sfericità verificata con il test di Mauchly $\chi^2(2) = 1.904$, $p = .386$, Greenhouse-Geisser $\varepsilon = 0.872$).

[20] Per tutti i confronti multipli post-hoc è stata applicata la correzione di Bonferroni con $\alpha = 0.017$.

nelle sezioni B1 e B2, solamente accennato per i candidati di livello B2. Infatti, in base alle analisi, i candidati di livello C conoscono in maniera eguale le combinazioni testate nelle sezioni B1 e B2 ($p = 0.43$, con $\alpha = 0.017$). In aggiunta, non è riscontrabile una differenza significativa neanche tra i punteggi delle sezioni B2 e C, sebbene per molto poco ($p = 0.02$, con $\alpha = 0.017$), indice del fatto che i candidati di livello C dimostrano in media una solida competenza trasversale sulle combinazioni di tutte le sezioni.

Grafico 3. *Punteggi medi totalizzati da candidati di livello QCER diverso in ogni sezione del COLL-IT*



6.3 *Analisi degli item*

Per rispondere alla domanda di ricerca 4, è stata svolta un'analisi degli item, che prevede il calcolo dell'*indice di facilità* (IF) e dell'*indice di discriminazione* (ID) degli item (quest'ultimo misurato come *corrected item-total correlation*, CITC)[21]. Da un punto di vista complessivo, l'analisi degli item indica un IF medio di 0.71 (DS 0.19) e un ID medio di 0.39 (DS 0.14). Prendendo in considerazione le tre sezioni del COLL-IT, il gruppo di item relativi al livello B1 riporta un IF medio di 0.76 (DS 0.17) e un ID medio di 0.37 (DS 0.13), al livello B2 un IF medio di 0.71 (DS 0.18) e ID medio di 0.4 (DS 0.12) e al livello C un IF medio di 0.66 (DS 0.21) e ID medio di 0.38 (DS 0.14)[22].

La Tabella 4 presenta un riassunto dell'analisi degli item, in cui le soglie di ID sono state in parte mutuate dalla proposta di Ebel (1979) e Popham (2000), e poi adattate in virtù del valore ID soglia adottato di 0.25 (Henning, 1987; Green, 2013).

Tabella 4. *Schema dell'analisi degli item suddivisi per livelli di difficoltà e discriminazione*

ITEM DEL TEST	**Difficili (IF < 0.20)**	**Medi (IF tra 0.20 e 0.80)**	**Facili (IF > 0.80)**
Scarsi (ID < 0.19)		68	5, 8, 10, 17, 18, 53, 75
Marginali (ID tra 0.19 e 0.25)	69	85, 95	25, 33, 58

[21] Per le analisi dettagliate di ogni item, si veda la Sezione 1 dell'Appendice.
[22] Per ulteriori dettagli, fare riferimento alla Sezione 1 dell'Appendice.

Accettabili (ID tra 0.25 e 0.29)		19, 36, 38, 72, 78, 99	35, 44, 79
Ragionevolmente buoni (ID tra 0.29 e 0.39)		6, 11, 12, 16, 24, 37, 42, 47, 51, 54, 56, 60, 81, 87, 93, 94	15, 21, 28, 40, 41, 48, 50, 55, 74, 91
Molto buoni (ID > 0.39)		2, 3, 13, 14, 20, 22, 26, 27, 30, 31, 34, 39, 45, 46, 49, 52, 61, 63, 65, 66, 67, 70, 71, 73, 82, 84, 86, 89, 90, 92	1, 4, 7, 9, 23, 29, 32, 43, 57, 59, 62, 64, 76, 77, 80, 83, 88, 96, 97, 98

In generale, si può dire che diversi item presentano valori di IF alti, ovvero oltre il valore soglia di 0.80 (42 item su 99, il 42,4% del totale), indice del fatto che i candidati coinvolti nella somministrazione hanno dimostrato nel complesso una buona conoscenza delle collocazioni bersaglio. In particolare, oltre il valore di 0.80 si trovano 17 item nella sezione B1 (51,5% dei 33 item), 14 in quella B2 (42,4%) e 11 in quella C (33,3%). Un solo item in tutto il test (item 69), relativo alla sezione C, ha riportato un IF inferiore alla soglia minima di 0.20, risultando complessivamente molto difficile per i candidati cui è stato somministrato il COLL-IT.

Per quanto concerne i valori di ID, 76 su 99 mostrano livelli di discriminazione ragionevolmente buoni o molto buoni, e altri 9 possono ritenersi accettabili, per un totale che ammonta all'85,9%. Nonostante ciò, i restanti 14 item risultano sotto la soglia ID minima di 0.25. Di questi, sette si trovano nella sezione B1 (item 5, 8, 10, 17, 18, 25, 33), due nella sezione B2 (item 53, 58) e cinque nella sezione C (item 68, 69, 75, 85, 95). Per alcuni di questi item è stata eseguita un'analisi dei distrattori, seguendo il metodo descritto da Allen e Yen (2002: 123). L'analisi quantifica la frequenza e la proporzione in percentuale delle risposte per ciascun item da parte dei partecipanti alla somministrazione, suddividendoli in tre sezioni omogenee: *candidati che hanno totalizzato un punteggio complessivo inferiore* (sezione punteggi bassi o SPB), *candidati con punteggio complessivo medio* (sezione punteggi medi o SPM), *candidati con punteggio complessivo alto* (sezione punteggi alti o SPA)[23].

In 8 casi su 14 (item 5, 8, 10, 17, 25, 33, 53, 58) l'IF risulta oltre il valore di 0.80 e la ragione della scarsa capacità di discriminazione degli item risiede nel fatto che sia i candidati più competenti, sia quelli medi, sia quelli meno competenti riescono a rispondere correttamente al quesito, in maniera difficilmente distinguibile. A mo' di esempio, si illustra la situazione dell'item 5 (IF 0.990; ID 0.145).

ITEM 5	Avete _____ la doccia o il bagno questa mattina?				
OPZIONI	dato	fatto*	avuto	creato	non lo so

La risposta esatta, contrassegnata nell'esempio da un asterisco (*), ovverosia in questo caso *fatto*, è stata selezionata dal 97% dei candidati della sezione punteggi complessivi bassi (SPB), e dal 100% dei candidati della sezione punteggi complessivi medi (SPM) e della sezione punteggi complessivi alti (SPA). Ne consegue che le opzioni distraenti proposte non hanno svolto adeguatamente la loro funzione.

Nel solo caso dell'item 69 (IF 0.165; ID 0.211) si verifica esattamente l'opposto rispetto a quanto descritto finora: il quesito è risultato troppo difficile sia per candidati SPA (risposte esatte per il 15% del gruppo) sia per quelli SPB (6%). L'opzione corretta (*stilato*) è stata selezionata da una proporzione maggiore di candidati SPM (29%), rispetto ai

[23] Si vedano i dettagli dell'analisi dei distrattori degli item nella sezione 2 dell'Appendice.

candidati di SPA. Questi ultimi hanno invece scelto il distrattore *svolto* in proporzioni maggiori (33%) rispetto agli altri gruppi (SPB: 6%; SPM: 12%). Inoltre, gran parte dei candidati di tutte le sezioni (41 candidati su 103) ha selezionato l'opzione "non lo so" (SPB: 52%; SPM: 37%; SPA 30%).

ITEM 69	È il quadro emerso dalle indagini dell'Agenzia Europea per l'ambiente che ha _____ la classifica delle aree più inquinate d'Europa.
OPZIONI	stilato* avuto filato svolto non lo so

L'item 18 presenta un IF alto (0.777) e un ID molto basso (0.1). L'opzione corretta (*lavare*) è stata selezionata dal 71% di SPB, dal 74% di SPM e dall'88% di SPA. Il distrattore più performante (*pulire*) ha attirato il 23% delle risposte di SPB, il 26% di SPM e il 12% di SPA. Gli altri distrattori sono risultati pressoché ininfluenti per tutti i livelli. Di fatto, l'opzione corretta è stata selezionata da un'ampia maggioranza di candidati. L'unica altra opzione distraente sembra non riuscire a discriminare con certezza tra i partecipanti SPB e SPM.

ITEM 18	I dentisti raccomandano di _____ i denti dopo ogni pasto e comunque due volte al giorno.
OPZIONI	lavare* pettinare pulire fare non lo so

L'item 68 presenta un IF tendente al valore ideale (0.553) ma un ID pressoché nullo (0.005). L'analisi dei distrattori mostra che l'opzione corretta (*conseguito*) è stata selezionata in misura maggiore dai candidati di SPB (54%) e SPM (66%) rispetto a quelli SPA, che l'hanno scelta nel 45% dei casi. Il distrattore che ha attirato maggiormente i candidati SPA è stato *raggiunto*, scelto dal 42% del gruppo, mentre i candidati SPB e SPM hanno optato per il distrattore *fatto*, rispettivamente per il 14% e per il 20%. Tali elementi contribuiscono ad abbassare il valore ID dell'item.

ITEM 68	Nel luglio 2015 ho _____ la laurea in Chimica e tecnologie farmaceutiche.
OPZIONI	conseguito* fatto raggiunto sollevato non lo so

L'item 75 ha un valore IF molto alto (0.845) e un ID molto basso (0.056). L'opzione corretta (*riduce*) è stata selezionata rispettivamente dall'89% di SPB, dal 69% di SPM e dal 97% di SPA. L'opzione distraente più selezionata è stata *abbassa*, che è stata scelta dal 6% di SPB, dal 26% di SPM e dallo 0% di SPA. La scarsa capacità discriminativa dipende dunque dal fatto che vi è stata una proporzione maggiore di candidati SPB che ha risposto correttamente al quesito rispetto ai candidati SPM, i quali, al contempo, sono stati attirati da un distrattore in maniera nettamente superiore (+20%) rispetto a SPB.

ITEM 75	La tecnologia _____ le distanze, è il mezzo che consente di superare quelle barriere che anni fa sembravano insormontabili.
OPZIONI	fa abbassa dà riduce* non lo so

L'item 85 presenta un IF di 0.786 e un ID pari a 0.241. L'opzione corretta (*tratta*) è stata selezionata dal 66% di SPB, 86% di SPM e 85% di SPA. Il distrattore più funzionale (*lavora*) ha attirato il 17% di SPB, il 6% di SPM e il 3% di SPA. L'opzione distraente *mette*, tuttavia, è stata scelta dal 6% di SPB, dal 3% di SPM e dal 9% di SPA. Questi elementi

mostrano che l'ID dell'item 85 è basso perché l'opzione corretta è stata selezionata in egual misura dai candidati di SPM e SPA, mentre il distrattore *mette*, invece, ha attirato una proporzione maggiore di candidati di SPA rispetto a SPB e SPM.

ITEM 85	L'autore _____ l'argomento con esempi chiari e a tratti divertenti.
OPZIONI	tratta* mette alza lavora non lo so

L'item 95 presenta valori di IF abbastanza alti (0.699) e un ID pari a 0.216. L'opzione corretta (*trovare*) è stata selezionata dai candidati di tutte le sezioni del campione in maniera crescente, rispetto al livello della sezione, ma simile (60%, 69% e 82%), mentre il distrattore *incontrare* ha attirato una proporzione maggiore di candidati SPM (17%) rispetto a quelli SPB (9%) e SPA (6%).

ITEM 95	È indispensabile _____ un compromesso tra costi e prestazioni.
OPZIONI	trovare* muovere dare incontrare non lo so

In linea di massima, si può affermare che il test COLL-IT si compone di item con discreti livelli medi di discriminazione. Tuttavia, nell'ottica di un processo di affinamento e miglioramento dello strumento, gli elementi emersi dall'analisi degli item e dall'analisi dei distrattori inducono a valutare un'eventuale revisione degli item che si sono dimostrati scarsamente performanti.

In aggiunta, in vista di future somministrazioni si può riflettere sull'opportunità di rimuovere alcuni item che si sono rivelati particolarmente difficili o facili, oltre che deficitari dal punto di vista della capacità discriminativa (per es. l'item 5 o l'item 69) per il gruppo di candidati bersaglio.

7. Conclusioni

In questo contributo è stato proposto il COLL-IT, un test volto a valutare la competenza collocazionale da parte di apprendenti di italiano L2/LS, ed è stato condiviso uno studio preliminare sul funzionamento dello strumento di verifica. L'obiettivo principale dello studio presentato era quello di indagare le prestazioni degli item del test, l'affidabilità dei punteggi, di operare inferenze sui punteggi e di valutare alcune prime evidenze di validità del costrutto di competenza collocazionale.

In § 2 è stato riproposto uno dei modelli teorici più celebri sul lessico e un suo successivo adattamento alle combinazioni lessicali. Inoltre, sono stati descritti alcuni test di valutazione della competenza collocazionale per apprendenti di inglese L2/LS. In § 3 è stato definito il costrutto alla base del presente studio e sono stati descritti il formato e il quadro metodologico sul quale si basa il COLL-IT. A seguire, in § 4 sono state riportate le attività di somministrazione del test e in § 5 sono state illustrate le domande di ricerca dello studio e la metodologia di indagine. Infine, in § 6 sono state riportate le analisi statistiche e la discussione dei risultati.

I risultati delle analisi si sono rivelati incoraggianti e nel complesso in linea con le aspettative. In primo luogo, la stima dell'indice di consistenza interna ha evidenziato un livello eccellente di affidabilità dei punteggi. In secondo luogo, le analisi statistiche hanno fornito prime evidenze riguardo alla bontà del costrutto interno di competenza collocazionale del COLL-IT, che è fondato sulla selezione delle collocazioni target a partire da un sillabo di combinazioni lessicali collegato ai livelli QCER. In terzo luogo,

sono state ottenute alcune evidenze preliminari sulla corrispondenza tra i punteggi del test e il livello QCER dei candidati. Da ultimo, l'analisi degli item ha fornito esiti positivi e ha fatto scaturire riflessioni costruttive sul funzionamento di alcuni item, sui quali occorrerà ritornare per eventuali modifiche e migliorie in vista di future somministrazioni.

Uno dei limiti principali del quadro teorico e metodologico della CTT è che i punteggi osservati sono validi per il campione di candidati cui è stato somministrato il test risultando, perciò, difficilmente generalizzabili con certezza su una popolazione più estesa, specie in assenza di ulteriori evidenze empiriche. Il coinvolgimento di un numero più cospicuo di candidati in future somministrazioni, inoltre, potrebbe favorire l'impiego di metodi di analisi degli item e dei punteggi più avanzati, come per esempio l'utilizzo del modello di Rasch o di altri modelli riconducibili alla *Item Response Theory*.

Un altro limite del presente studio è rappresentato dalla mancanza di evidenze di validità concorrente. Con lo scopo di valutare le correlazioni tra i punteggi su test simili, occorrerebbe somministrare il COLL-IT contestualmente a un altro (o ad altri) test, che si proponga di misurare il medesimo costrutto o delle competenze affini. In questo modo, si potrebbero ottenere ulteriori evidenze utili ad avvalorare la validità del test.

Infine, con la prospettiva di approfondire la comprensione di un costrutto complesso quale è la competenza collocazionale e il suo sviluppo, sarebbe utile svolgere ulteriori approfondimenti in merito ad alcune proprietà delle collocazioni target presenti negli item del COLL-IT. Per esempio, si potrebbe considerare il ruolo della congruenza tra collocazioni target e combinazioni lessicali presenti nella lingua madre (o altre lingue conosciute) di gruppi di apprendenti di diverse L1, indagando gli effetti di tale fattore sui punteggi ottenuti a diversi livelli di competenza linguistico-comunicativa.

Nonostante i limiti delineati, la creazione e la condivisione di uno strumento di verifica della competenza collocazionale di apprendenti di italiano, realizzato a partire da riflessioni documentate e da una solida metodologia, si propongono di offrire un contributo nel panorama della ricerca in ambito valutativo L2/LS.

RIFERIMENTI BIBLIOGRAFICI

Allen M. J., Yen W. M. (2002), *Introduction to Measurement Theory*, Waveland Press, Prospect Heights, Long Grove (IL-US).

ALTE Members (1998), *Multilingual Glossary of Language Testing Terms*, Cambridge University Press Cambridge, Cambridge.

Bachman L. F. (1990), *Fundamental Considerations in Language Testing*, Oxford University Press, Oxford.

Bachman L. F. (2004), *Statistical Analyses for Language Assessment*, University of California, Los Angeles: https://doi.org/10.1017/cbo9780511667350.

Bachman L. F., Palmer A. S. (2010), *Language assessment in practice*, Oxford University Press, Oxford.

Bardel C., Lindqvist C., Laufer B. (eds.) (2013), *L2 Vocabulary Acquisition, Knowledge and Use. New Perspectives on Assessment and Corpus Analysis*, Eurosla Monograph Series 2: http://www.eurosla.org/monographs/EM02/TOC.pdf.

Barni M. (2023), *Valutare le competenze nelle L2. Teorie, metodi, strumenti, politiche linguistiche*, Carocci, Roma.

Berišić Antić D. (2015), "Le Collocazioni italiane nell'insegnamento dell'italiano come L2", in *Strani jezici*, 44, 4, pp. 260-278.

Bestgen Y., Granger S. (2014), "Quantifying the development of phraseological competence in L2 English writing: An automated approach", in *Journal of Second Language Writing*, 26, pp. 28-41: https://doi.org/10.1016/j.jslw.2014.09.004.

Bini M., Pernas A., Pernas P. (2007), "Apprendimento e insegnamento collocazioni dell'italiano. Con i NUNC più facile", in Barbera M., Corino E., Onesti C. (a cura di), *Corpora e linguistica in rete*, Guerra Edizioni, Perugia, pp. 323-333.

Biskup D. (1992), "L1 influence on learners'renderings of English collocations. A Polish/German empirical study", in Arnaud P. J. L., Béjoint H. (eds.), *Vocabulary and Applied Linguistics*, Macmillan, London, pp. 85-93: https://doi.org/10.1007/978-1-349-12396-4_8.

Canty A., Ripley B. D. (2022), *boot: Bootstrap R (S-Plus) Functions*, R package version 1.3-28.1.

Chapelle C. A. (2012), "Validity argument for language assessment: The framework is simple…", in *Language Testing*, 29, 1, pp. 19-27. https://doi.org/10.1177/0265532211417211.

Chapelle C. A. (2021), *Argument based validation in testing and assessment*, Sage Publishing, Thousand Oaks (CA-US).

Consiglio d'Europa (2002), *Quadro comune europeo di riferimento per le lingue: apprendimento, insegnamento, valutazione*, La Nuova Italia-Oxford, Firenze.

Cronbach L. (1951), "Coefficient alpha and the internal structure of tests", in *Psychometrika*, 16, pp. 292-334: https://doi.org/10.1007/BF02310555.

Daller H., Milton J., Treffers-Daller J. (eds.) (2007), *Modelling and Assessing Vocabulary Knowledge*, Cambridge University Press, Cambridge: https://doi.org/10.1017/cbo9780511667268.

DeVellis R. F., Thorpe C. T. (2022[5]), *Scale development: Theory and applications*, SAGE publications, Thousand Oaks (CA-US).

DeVellis R. F. (1991), *Scale Development*, Sage Publications, Newbury Park (NJ-US).

Dinno A. (2017). *dunn.test: Dunn's Test of Multiple Comparisons Using Rank Sums*, R package version 1.3.5.

Dóczi B., Kormos J. (2016), *Longitudinal Development in Vocabulary Knowledge and Lexical Organization*, Oxford University Press, Oxford: https://doi.org/10.1111/ijal.12179.

Dunn T. J., Baguley T., Brunsden V. (2014), "From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation", in *British Journal of Psychology*, 105, pp. 399-412: https://doi.org/10.1111/bjop.12046.

Ebel R.L. (1979), *Essentials of Educational Measurement*, Prentice Hall, New Jersey.

Faerch C., Haastrup K., Phillipson R. (1984), *Learner Language and Language Learning*, Multilingual Matters, Clevedon (UK).

Farghal M., Obiedat H. (1995), "Collocations: A neglected variable in EFL", in *International Journal of Applied Linguistics*, 28, 4, pp. 313-331: https://doi.org/10.1515/iral.1995.33.4.315.

Fox J., Weisberg S. (2019[3]), *An R Companion to Applied Regression*, Sage Publishing, Thousand Oaks (CA-US).

Gallina F. (2018), "Studenti internazionali in mobilità: la questione del lessico della conoscenza in italiano L2", in Coonan C. M., Bier A, Ballarin E. (a cura di), *La didattica delle lingue nel nuovo millennio Le sfide dell'internazionalizzazione*, pp. 323-339: https://doi.org/10.14277/6969-227-7/SR-13-20.

Gallina F. (2022), *Osservare e valutare la competenza lessicale in italiano L2*, Pacini Editore, Pisa.

Gitsaki C. (1999), *Second Language Lexical Acquisition: A Study of the development of collocational knowledge*, International Scholars Publications, San Francisco.

Green R. (2013), *Statistical analyses for language testers*, Palgrave Macmillan, Basingstoke (UK): https://doi.org/10.1057/9781137018298.

Gyllstad H. (2005), "Words that go together well: Developing test formats for measuring learner knowledge of English collocations", in Heinat F., Klingval E. (eds.), T*he Department of English in Lund: Working papers in linguistics*, Lund University, Lund, pp. 1-31.

Gyllstad H. (2007), *Testing English collocations: Developing receptive tests for use with advanced Swedish learners* [tesi di dottorato], Lund University, Lund.

Gyllstad H. (2009), "Designing and evaluating tests of receptive collocation knowledge: COLLEX and COLLOMATCH", in Barfield A., Gyllstad H. (eds.), *Researching collocations in another language*, Palgrave Macmillan, Basingstoke (UK), pp. 153-170: https://doi.org/10.1057/9780230245327_12.

Hargreaves P. (2000), "How important is collocation in testing the learner's language proficiency?", in Lewis M. (ed.), *Teaching Collocation: Further Developments in the Lexical Approach*, Language Teaching Publications, Hove (UK), pp. 205-223.

Hasselgren A. (1994), "Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary", in *International Journal of Applied Linguistics*, 4, pp. 237-258: https://doi.org/10.1111/j.1473-4192.1994.tb00065.x.

Henning G. (1987), *A guide to language testing: development, evaluation, and research*, Newbury House, New York.

Henriksen B. (2013), "Research on L2 learners' collocational competence and development – A progress report", in Bardel C., Lindqvist C., Laufer B. (eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis*, EuroSLA Monographs Series, Amsterdam, pp. 29-56.

Jezek E. (2005), *Lessico. Classi di parole, strutture, combinazioni*, il Mulino, Bologna.

Kane M. (2012), "Validating score interpretations and uses", in *Language Testing,* 29, 1, pp. 3-17: https://doi.org/10.1027/1614-2241/a000036.

Kelley K., Cheng Y. (2012), "Estimation of and confidence interval formation for reliability coefficients of homogeneous measurement instruments", in *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 8, 2, pp. 39-50: https://doi.org/10.1027/1614-2241/a000036.

Kilgarriff A., Husák M., McAdam K., Rundell M., Rychlý P. (2008), "GDEX: Automatically finding good dictionary examples in a corpus", in *Proceedings of the XIII EURALEX international* congress, Vol. 1, Universitat Pompeu Fabra, Barcelona, pp. 425-432.

La Russa F., D'Alessio V., Suadoni A. (2023), "Designing a Corpus-Based Syllabus of Italian Collocations: Criteria, Methods and Procedures", in *Revue roumaine de linguistique*, pp. 377-389: https://doi.org/10.54103/2037-3597/20398.

Laufer B., Waldman T. (2011), "Verb-Noun Collocations in Second Language Writing: A Corpus Analysis of Learners' English", in *Language Learning*, 61, 2, pp. 647-672: https://doi.org/10.1111/j.1467-9922.2010.00621.x.

Lawrence M. (2016). *ez: Easy Analysis and Visualization of Factorial Experiments*, R package version 4.4-0.

Lewis M. (1993), *The lexical approach*, Language Teaching Publications, Hove (UK).

Lewis M. (ed.)(2000), *Teaching collocation: Further developments in the lexical approach*, Language Teaching Publications.

McNamara T. (2000), *Language Testing*, Oxford University Press, Oxford.

Meara P., Miralpeix I. (2017), *Tools for Researching Vocabulary*, Multilingual Matters, Bristol: https://doi.org/10.21832/9781783096473.

Mel'Cuk I. A., Wanner L. (1994), "Lexical co-occurrence and lexical inheritance. Emotion lexemes in German: A lexicographic case study", in *Lexikos*, 4, 4, pp. 86-161.

Messick S. (1989), "Validity", in Linn R. L. (ed.) *Educational measurement*, American Council on education and Macmillan, New York, pp. 13-104.

Milton J. (2009), *Measuring Second Language Vocabulary Acquisition*, Multilingual Matters, Bristol: https://doi.org/10.21832/9781847692092.

Nation P., Webb S. (2011), *Researching and Analysing Vocabular*, Heinle, Cengage Learning, Boston.

Nattinger J., DeCarrico J. (1992), *Lexical Phrases and Language Teaching*, Oxford University Press, Oxford.

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

Omidian T., Siyanova-Chanturia A., Spina S. (2021), "Development of formulaic knowledge in learner writing: A longitudinal perspective", in Granger S. (ed.). *Perspectives on the Second Language Phrasicon: The View from Learner Corpora*, Multilingual Matters, Bristol, pp. 178-206: https://doi.org/10.21832/9781788924863-009.

Pallone S. (2023), *Apprendimento di costruzioni a verbo supporto dell'italiano da parte di studenti lusofoni: un confronto tra contesto di immersione e di lingua straniera.* Tesi di laurea magistrale sostenuta presso l'Università degli Studi Roma Tre.

Pérez Serrano M. (2017), "Evaluación de la competencia colocacional: Una revisión bibliográfica", in *Boletín de ASELE*, 57, pp. 69-78.

Popham W. J. (2000[3]), *Modern Educational Measurement*, Allyn & Bacon, Boston.

Read J. (2000), *Assessing Vocabulary*, Cambridge University Press, Cambridge: https://doi.org/10.1017/CBO9780511732942.

Revelle W. (2017), *Psych: Procedures for psychological, psychometric, and personality research*: https://cran.r-project.org/package=psych (R package version 1.8.12),

Revier R. (2009), "Evaluating a New Test on Whole English Collocations", in Barfield A., Gyllstad H. (eds.), *Researching collocations in another language*, Palgrave Macmillan, Basingstoke (UK), pp. 125-138: https://doi.org/10.1057/9780230245327_10.

RStudio Team (2021), *RStudio: Integrated Development Environment for* R., RStudio, PBC, Boston: http://www.rstudio.com/.

Schmitt N. (2010), *Researching Vocabulary. A Vocabulary Research Manual*, Palgrave Macmillan, Basingstoke (UK): https://doi.org/10.1057/9780230293977.

Sijtsma K. (2009), "On the use, the misuse, and the very limited usefulness of Cronbach's alpha", in *Psychometrika*, 74, pp. 107-120: https://doi.org/10.1007/s11336-008-9101-0.

Simone R. (1996), *Fondamenti di linguistica*, Laterza, Roma-Bari.

Sinclair J. (1991), *Corpus, concordance, collocation*, Oxford University Press, Oxford.

Siyanova-Chanturia A. (2015), "Collocation in beginner learner writing: A longitudinal study", in *System*, 53, pp. 148-160: https://doi.org/10.1016/j.system.2015.07.003.

Siyanova-Chanturia A., Spina S. (2020), "Multi-word expressions in second language writing: A large-scale longitudinal learner corpus study", in *Language learning*, 70, 2, pp. 420-463: https://doi.org/10.1111/lang.12383.

Spina S. (2014), "Il Perugia Corpus: una risorsa di riferimento per l'italiano: composizione, annotazione e valutazione", in Basili R., Lenci A., Magnini B. (a cura di), *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014*, Pisa University Press, Pisa, pp. 354-359: https://doi.org/10.12871/clicit2014168.

Spina S. (2016), "Learner corpus research and phraseology in Italian as a second language: The case of the DICI-A, a learner dictionary of Italian collocations", in Sanromán Vilas B. (ed.), *Collocations Cross-Linguistically. Corpora, Dictionaries and Language Teaching*, Mémoires de la Société Néophilologique de Helsink, Uusfilologinen yhdistys ry, Helsinki, pp. 219-244.

Spina S. (2018), "Lo sviluppo longitudinale della fraseologia in apprendenti cinesi di italiano L2. Uno studio preliminare su alcune categorie di errori", in *Ricognizioni. Rivista di lingue, letterature e culture moderne*, 10, 5, pp. 97-119.

Spina S. (2019), "The development of phraseological errors in Chinese learner Italian: A longitudinal study", in Abel A., Glaznieks A., Lyding V., Nicolas L. (eds.), *Widening the scope of learner corpus research. Selected papers from the fourth Learner Corpus Research Conference*, Presses Universitaires de Louvain, Louvain, pp. 95-119.

Spina S. (2022), "The effect of time and dimensions of collocational relationship on phraseological accuracy: a study on Chinese learners of Italian", in Leńko-Szymańska A., Götz S. (eds.), *Complexity, Accuracy & Fluency in Learner Corpus Research*, John Benjamins, Amsterdam, pp. 181-207: https://doi.org/10.1075/scl.104.08spi.

Spina S., Fioravanti I., Forti L., Santucci V., Serra A., Zanda F. (2022), "Il corpus CELI: una nuova risorsa per studiare l'acquisizione dell'italiano L2", in *Italiano LinguaDue*, 14, 1, pp. 116-138: https://doi.org/10.54103/2037-3597/18161.

Spinelli B., Parizzi F. (2010), *Profilo della lingua italiana. Livelli di riferimento del QCER A1, A2, B1, B2*, La Nuova Italia, Firenze.

Suadoni A. (2020), "Acquisition of Italian multi-word verbs and collocations by Spanish-speaking learners", in *Rivista di psicolinguistica applicata*, 20, 2, pp. 101-120: https://doi.org/10.19272/202007702007.

Wang Y. (2016), *The Idiom Principle and L1 Influence. A contrastive learner-corpus study of delexical verb+noun collocations*, John Benjamins, Amsterdam: https://doi.org/10.1075/scl.77.

Weir C. J. (2005), *Language Testing and Validation: An Evidence-Based Approach*, Palgrave Macmillan, Basingstoke (UK): https://doi.org/10.1057/9780230514577.

Willse J. T. (2018), *CTT: Classical test theory functions* (Computer software manual): https://cran.r-project.org/package=CTT (R package version 2.3.3).

Yoon H. J. (2016), "Association strength of verb-noun combinations in experienced NS and less experienced NNS writing: Longitudinal and cross-sectional findings", in *Journal of Second Language Writing*, 34, pp. 42-57: https://doi.org/10.1016/j.jslw.2016.11.001.

APPENDICE

Sezione 1 - Analisi degli item suddivisi per sezioni del test, con valori IF e ID (CITC).

	SEZIONE TEST B1				SEZIONE TEST B2				SEZIONE TEST C		
N° ITEM	**Collocazione target dell'item**	**IF**	**ID**	**N° ITEM**	**Collocazione target dell'item**	**IF**	**ID**	**N° ITEM**	**Collocazione target dell'item**	**IF**	**ID**
1	avere bisogno	0.883	0.393	**34**	vivere situazione	0.621	0.427	**67**	farsi carico	0.563	0.427
2	fare lavoro	0.796	0.426	**35**	lasciare lavoro	0.864	0.275	**68**	conseguire laurea	0.553	0.003
3	accettare invito	0.864	0.389	**36**	dare svolta	0.330	0.287	**69**	stilare classifica	0.165	0.190
4	perdere lavoro	0.854	0.426	**37**	girare mondo	0.786	0.361	**70**	fare fronte	0.621	0.471
5	fare doccia	0.990	0.145	**38**	dare giudizio	0.282	0.277	**71**	adottare provvedimento	0.515	0.433

6	seguire corso	0.427	0.351	**39**	fare salto	0.505	0.420	**72**	assegnare premio	0.757	0.282
7	avere voglia	0.845	0.561	**40**	sentire presenza	0.825	0.358	**73**	dare sguardo	0.524	0.512
8	alzare mano	0.883	0.000	**41**	creare ambiente	0.816	0.334	**74**	allungare vita	0.854	0.332
9	chiedere aiuto	0.864	0.481	**42**	considerare fatto	0.748	0.384	**75**	ridurre distanza	0.845	0.056
10	cantare canzone	0.981	0.148	**43**	offrire servizio	0.942	0.428	**76**	adottare politica	0.845	0.439
11	fare stage	0.534	0.345	**44**	aumentare prezzo	0.913	0.262	**77**	contrarre malattia	0.786	0.463
12	prendere parte	0.592	0.381	**45**	evitare spreco	0.650	0.486	**78**	dare peso	0.777	0.281
13	chiedere favore	0.748	0.576	**46**	raggiungere scopo	0.505	0.607	**79**	dire parola	0.922	0.278
14	fare acquisto	0.680	0.481	**47**	trovare appoggio	0.670	0.295	**80**	negare fatto	0.942	0.426
15	festeggiare compleanno	0.971	0.294	**48**	perdere tempo	0.893	0.301	**81**	recare danno	0.282	0.376
16	chiedere permesso	0.786	0.574	**49**	svolgere attività	0.563	0.661	**82**	vedere ombra	0.573	0.400
17	dare definizione	0.816	0.069	**50**	rispettare legge	0.883	0.361	**83**	rispettare scelta	0.816	0.531
18	lavare dente	0.777	0.076	**51**	mantenere contatto	0.680	0.293	**84**	subire violenza	0.650	0.580
19	sostenere esame	0.262	0.282	**52**	tradire fiducia	0.427	0.598	**85**	trattare argomento	0.786	0.241
SEZIONE TEST B1				**SEZIONE TEST B2**				**SEZIONE TEST C**			
N° ITEM	**Collocazione target dell'item**	**IF**	**ID**	**N° ITEM**	**Collocazione target dell'item**	**IF**	**ID**	**N° ITEM**	**Collocazione target dell'item**	**IF**	**ID**
20	dare parere	0.563	0.411	**53**	apprendere lingua	0.806	0.145	**86**	prendere respiro	0.767	0.522
21	finire scuola	0.913	0.308	**54**	dare coraggio	0.680	0.328	**87**	prendere sopravvento	0.340	0.366
22	cercare aiuto	0.767	0.463	**55**	perdere vita	0.854	0.381	**88**	dare continuità	0.806	0.523
23	perdere treno	0.816	0.543	**56**	dare spiegazione	0.767	0.385	**89**	porre quesito	0.233	0.539
24	trascorrere vacanza	0.505	0.311	**57**	aprire mente	0.893	0.551	**90**	stabilire rapporto	0.592	0.545
25	fare discorso	0.913	0.242	**58**	raccogliere informazione	0.864	0.194	**91**	trasmettere messaggio	0.825	0.358
26	dare voto	0.689	0.449	**59**	spendere soldo	0.806	0.488	**92**	soddisfare curiosità	0.505	0.493
27	fare schifo	0.670	0.615	**60**	superare difficoltà	0.796	0.335	**93**	prendere misura	0.680	0.363

28	attraversare strada	0.864	0.361	**61**	ridare speranza	0.427	0.487	**94**	allargare orizzonte	0.311	0.343
29	cercare lavoro	0.816	0.420	**62**	avere intenzione	0.874	0.420	**95**	trovare compromesso	0.699	0.191
30	fare giro	0.583	0.447	**63**	accettare idea	0.680	0.470	**96**	danneggiare salute	0.893	0.415
31	accendere televisione	0.534	0.569	**64**	trovare ispirazione	0.903	0.512	**97**	imporre modello	0.806	0.576
32	avere paura	0.806	0.594	**65**	accettare situazione	0.680	0.451	**98**	avere presentimento	0.893	0.482
33	prendere taxi	0.961	0.193	**66**	abbassare costo	0.592	0.613	**99**	mantenere promessa	0.777	0.267
B1	**MEDIA SEZIONE (DS)**	**0.76 (0.17)**	**0.37 (0.13)**	**B2**	**MEDIA SEZIONE (DS)**	**0.71 (0.18)**	**0.4 (0.12)**	**C**	**MEDIA SEZIONE (DS)**	**0.66 (0.21)**	**0.38 (0.14)**

Sezione 2 - Analisi dei distrattori degli item 5, 18, 68, 69, 75, 85 e 95.

ITEM 5 Avete _____ la doccia o il bagno questa mattina?
OPZIONI dato fatto* avuto creato non lo so

RISPOSTE	**SEZIONE PUNTEGGI BASSI (SPB)**		**SEZIONE PUNTEGGI MEDI (SPM)**		**SEZIONE PUNTEGGI ALTI (SPA)**	
	Frequenza	Proporzione	Frequenza	Proporzione	Frequenza	Proporzione
*fatto	34	0.971	35	1.000	33	1.000
dato	0	0.000	0	0.000	0	0.000
avuto	0	0.000	0	0.000	0	0.000
creato	0	0.000	0	0.000	0	0.000
non lo so	1	0.029	0	0.000	0	0.000

ITEM 18 I dentisti raccomandano di _____ i denti dopo ogni pasto e comunque due volte al giorno.
OPZIONI lavare* pettinare pulire fare non lo so

RISPOSTE	**SEZIONE PUNTEGGI BASSI (SPB)**		**SEZIONE PUNTEGGI MEDI (SPM)**		**SEZIONE PUNTEGGI ALTI (SPA)**	
	Frequenza	Proporzione	Frequenza	Proporzione	Frequenza	Proporzione
*lavare	25	0.714	26	0.743	29	0.879
pettinare	0	0.000	0	0.000	0	0.000
pulire	8	0.229	9	0.257	4	0.121
fare	0	0.000	0	0.000	0	0.000
non lo so	2	0.057	0	0.000	0	0.000

ITEM 68 Nel luglio 2015 ho _____ la laurea in Chimica e tecnologie farmaceutiche.
OPZIONI conseguito* fatto raggiunto sollevato non lo so

RISPOSTE	**SEZIONE PUNTEGGI BASSI (SPB)**		**SEZIONE PUNTEGGI MEDI (SPM)**		**SEZIONE PUNTEGGI ALTI (SPA)**	
	Frequenza	Proporzione	Frequenza	Proporzione	Frequenza	Proporzione
*conseguito	19	0.543	23	0.657	15	0.455
fatto	5	0.143	7	0.200	1	0.030
sollevato	1	0.029	1	0.029	1	0.030
raggiunto	1	0.029	4	0.114	14	0.424
non lo so	9	0.257	0	0.000	2	0.061

ITEM 69 È il quadro emerso dalle indagini dell'Agenzia Europea per l'ambiente che ha _____ la classifica delle aree più inquinate d'Europa.
OPZIONI stilato* avuto filato svolto non lo so

RISPOSTE	**SEZIONE PUNTEGGI BASSI (SPB)**		**SEZIONE PUNTEGGI MEDI (SPM)**		**SEZIONE PUNTEGGI ALTI (SPA)**	
	Frequenza	Proporzione	Frequenza	Proporzione	Frequenza	Proporzione
*stilato	2	0.057	10	0.286	5	0.152
avuto	11	0.314	7	0.200	4	0.121
filato	2	0.057	1	0.029	3	0.091
svolto	2	0.057	4	0.114	11	0.333
non lo so	18	0.514	13	0.371	10	0.303

ITEM 75 La tecnologia _____ le distanze, è il mezzo che consente di superare quelle barriere che anni fa sembravano insormontabili.
OPZIONI fa abbassa dà riduce* non lo so

RISPOSTE	**SEZIONE PUNTEGGI BASSI (SPB)**		**SEZIONE PUNTEGGI MEDI (SPM)**		**SEZIONE PUNTEGGI ALTI (SPA)**	
	Frequenza	Proporzione	Frequenza	Proporzione	Frequenza	Proporzione
*riduce	31	0.886	24	0.686	32	0.970
abbassa	2	0.057	9	0.257	0	0.000
fa	1	0.029	0	0.000	0	0.000

dà	0	0.171	0	0.000	0	0.000
non lo so	1	0.029	2	0.057	1	0.030

ITEM 85 L'autore _____ l'argomento con esempi chiari e a tratti divertenti.
OPZIONI tratta* mette alza lavora non lo so

RISPOSTE	**SEZIONE PUNTEGGI BASSI (SPB)**		**SEZIONE PUNTEGGI MEDI (SPM)**		**SEZIONE PUNTEGGI ALTI (SPA)**	
	Frequenza	Proporzione	Frequenza	Proporzione	Frequenza	Proporzione
*tratta	23	0.657	30	0.857	28	0.848
mette	2	0.057	1	0.029	3	0.091
alza	1	0.029	1	0.029	0	0.000
lavora	6	0.171	2	0.057	1	0.030
non lo so	3	0.086	1	0.029	1	0.030

ITEM 95 È indispensabile _____ un compromesso tra costi e prestazioni.
OPZIONI trovare* muovere dare incontrare non lo so

RISPOSTE	**SEZIONE PUNTEGGI BASSI (SPB)**		**SEZIONE PUNTEGGI MEDI (SPM)**		**SEZIONE PUNTEGGI ALTI (SPA)**	
	Frequenza	Proporzione	Frequenza	Proporzione	Frequenza	Proporzione
*trovare	21	0.600	24	0.686	27	0.818
muovere	0	0.000	1	0.029	0	0.030
dare	3	0.086	2	0.057	2	0.061
incontrare	3	0.086	6	0.171	2	0.061
non lo so	8	0.229	2	0.057	2	0.061