FEDERICA DEL BONO[1]

Aspetti pragmatici nella valutazione di testi scritti: uno studio sull'adeguatezza funzionale in italiano L2

Abstract

Functional adequacy, intended as an interpersonal and task-related construct (Kuiken - Vedder, 2017), is an important aspect of the pragmatic competence to consider when assessing L2 writing. According to Kuiken - Vedder (2017) functional adequacy can be assessed by means of a rating scale which splits up the construct into four dimensions: content, task requirements, comprehensibility, coherence and cohesion. This paper presents the results of a study set up to test the applicability of the rating scale in Italian L2 for non-expert raters – native speakers of Italian – in order to assess functional adequacy in L2 writing, on the basis of a narrative, a decision making and an instruction task.. The results confirm the possibility to use the rating scale for assessing functional adequacy and provides useful information concerning the applicability of the scale for different task types.

1. *Introduzione*

Il nuovo *companion volume* del Quadro Comune Europeo di Riferimento per le lingue (QCER) ha nuovamente sottolineato la distinzione tra competenza linguistica – intesa come uso della lingua relativo alle sue risorse e alla sua conoscenza in quanto sistema – e competenza pragmatica – intesa come uso reale della lingua nella costruzione dei testi (Council of Europe, 2018: 136). Per essere competenti in L2 non è sufficiente esserlo dal punto di vista linguistico, ma bisogna anche esserlo dal punto di vista pragmatico. Una componente importante della competenza pragmatica è la dimensione funzionale del linguaggio che, come affermano Kuiken e Vedder (2017), deve essere presa in considerazione quando si valuta una produzione scritta in L2. Come evidenziato da Kuiken e Vedder (2017), sebbene l'importanza degli aspetti funzionali della lingua sia affermata in diversi studi, quelli che si dedicano a studiarli sono numericamente inferiori rispetto a quelli che prendono in esame gli aspetti della complessità, dell'accuratezza e della fluenza (CAF) in produzioni scritte.

La scarsità di studi sulla dimensione funzionale del linguaggio è da attribuirsi principalmente alla mancanza di unanimità nel darne una definizione (Kuiken - Vedder, 2017: 322). Infatti, se si osservano gli studi dedicati all'argomento si nota

[1] Università degli Studi Roma Tre.

immediatamente come questi utilizzino diversi termini e definizioni per riferirsi alla dimensione funzionale del linguaggio. Ad esempio, Hismanoglu (2011) parla di *competenza comunicativa interculturale*, intesa come scelta di argomenti conversazionali appropriati, aperture e chiusure conversazionali, stereotipi e comunicazione non verbale. Al contrario, Bridgeman *et al.* (2011) parlano di *communicative effectiveness*, intesa in termini di comprensibilità del messaggio. Altri autori (Pallotti, 2009; Kuiken *et al.*, 2010; De Jong *et al.*, 2012a, 2012b), invece, parlano di *adeguatezza comunicativa*, intesa come riuscita esecuzione di un *task* (per una rassegna di studi sulla dimensione funzionale del linguaggio si veda Kuiken - Vedder, 2017).

La mancanza di una definizione unanime della dimensione funzionale del linguaggio ha portato a una conseguente mancata unanimità nella scelta degli strumenti per valutarla. Infatti, come affermano Kuiken *et al.* (2010) a differenza degli indici CAF (Pallotti, 2009), non esistono delle eguali misure oggettive per valutare la dimensione funzionale del linguaggio. Se si osservano gli studi dedicati alla valutazione di questa dimensione, emerge che questi si sono affidati a diversi criteri e strumenti per valutarla (per una rassegna di studi sulla valutazione della dimensione funzionale del linguaggio si veda Kuiken - Vedder, 2017).

Il presente studio fa riferimento alla dimensione funzionale del linguaggio adottando il termine *adeguatezza funzionale* (da qui in poi FA – dall'inglese *Functional Adequacy*) secondo la definizione che ne hanno dato Kuiken e Vedder (2017) e ha l'obiettivo di testare l'applicabilità della scala di valutazione della FA ideata dagli stessi su tre tipi di testi scritti: narrativi, argomentativi, regolativi.

2. *L'adeguatezza funzionale: che cos'è e come valutarla*

Kuiken e Vedder (2017) hanno definito la FA come un costrutto interpersonale – in quanto coinvolge due partecipanti (A e B) – relativo a un task e alla sua riuscita esecuzione (cfr. anche Vedder, 2016). Questa definizione è in linea con quanto già sostenuto da altri autori (Pallotti, 2009; Kuiken *et al.*, 2010; De Jong *et al.*, 2012a, 2012b). Ciò a cui, però, Kuiken e Vedder prestano particolare attenzione è il concetto che l'adeguatezza del messaggio prodotto da A e ricevuto da B varia a seconda del tipo di task che i partecipanti stanno svolgendo. È proprio questo nesso tra la riuscita esecuzione del task e l'adeguatezza del messaggio che ha portato gli autori a utilizzare il termine *funzionale*, invece che *comunicativa* (Kuiken - Vedder, 2017: 323).

Kuiken e Vedder (2017: 326) nel loro contributo hanno messo in luce la necessità di ideare un mezzo di valutazione della FA che avesse dei descrittori oggettivi, indipendenti dagli indici CAF e che fosse applicabile sia in L1 che in L2 da parte di valutatori esperti e non. Partendo da queste idee di base, Kuiken e Vedder hanno proposto una nuova scala di valutazione del costrutto. Tale scala ha avuto come ulteriore fondamento teorico le massime conversazionali di Grice (1975), in quanto l'adeguatezza del messaggio viene interpretata in termini di quantità, qualità, modo e relazione dello stesso (cfr. Kuiken - Vedder, 2017: 323). Per ideare la scala di va-

lutazione Kuiken e Vedder si sono affidati a due elementi: i descrittori del QCER (Consiglio d'Europa, 2002) e la scala di valutazione dell'adeguatezza comunicativa di De Jong *et al.* (2012a; 2012b). Nella scala di Kuiken - Vedder (2017) l'adeguatezza funzionale è scomposta in quattro dimensioni: contenuto, requisiti del tipo di task, comprensibilità, coerenza e coesione. Ogni dimensione è ispirata a una o più massime di Grice ed è rappresentata da una scala a sei punti con descrittori specifici per ogni punto. Il contenuto fa riferimento alle unità informative (idee) contenute nel testo, le quali devono essere efficaci e adeguate nel numero (massime di quantità, qualità e relazione). La dimensione dei requisiti del tipo di task fa riferimento al rispetto dei criteri specifici del task intesi in termini di registro, atti linguistici e genere testuale (massime di relazione e modo). La comprensibilità valuta quanto il testo sia comprensibile e quanto sforzo il lettore deve impiegare per comprenderlo (massima di modo). Infine, la coerenza e coesione fa riferimento alla presenza o meno di salti logici e di strumenti coesivi nel testo (massima di modo).

Kuiken - Vedder (2017) hanno testato l'applicabilità della scala sui testi argomentativi in italiano e olandese L2 e L1. Nel loro studio studenti universitari hanno dovuto produrre due testi argomentativi, valutati da valutatori non esperti, anch'essi studenti universitari e parlanti nativi della lingua in cui era scritto il testo. Prima della sessione di valutazione, i valutatori sono stati sottoposti a due sessioni di *training* in compresenza per prendere familiarità con l'uso delle scale. Dai risultati dello studio è emerso che la scala è uno strumento affidabile per valutare la FA in testi argomentativi sia in L1 che in L2 e che il costrutto può essere valutato in termini di quattro dimensioni. Infine, è anche emerso che la scala può essere utilizzata sia da valutatori esperti sia da valutatori non esperti.

2.1. Studi sulle scale di valutazione dell'adeguatezza funzionale in italiano L2

Tra i suggerimenti sugli aspetti della scala che bisognerebbe ulteriormente studiare, Kuiken e Vedder (2017) hanno sottolineato la necessità di verificare l'applicabilità della stessa per valutare diversi tipi di testi scritti. Partendo da questo suggerimento, sono usciti nuovi studi che hanno testato l'applicabilità della scala in italiano L2 e L1 su diversi tipi di testi scritti.

Il primo studio è quello di Cortés Velásquez - Nuzzo (2017), nel quale è stata verificata l'applicabilità della scala su e-mail di tipo persuasivo prodotte da studenti universitari in italiano L1. I risultati hanno confermato la possibilità di applicare la scala per valutare testi in L1 da parte di valutatori non esperti. Ciò nonostante, dall'analisi dei dati è emerso uno scarso accordo assoluto tra i valutatori che secondo gli autori è dovuto a problemi di interpretazione dei descrittori.

Un secondo studio è quello di Faone e Pagliara (2017) nel quale si è testata l'applicabilità della scala su testi regolativi in italiano L2 prodotti da 15 studenti sinofoni al termine di un corso di italiano L2. Nello studio sono stati impiegati 3 valutatori esperti e 3 valutatori non esperti, i quali sono stati addestrati all'uso delle scale della FA, per un totale di 2 sessioni di *training* in compresenza. Dopo la fase di *training* i valutatori hanno valutato individualmente i testi regolativi prodotti dagli

studenti sinofoni. I risultati dello studio hanno confermato l'affidabilità della scala. Inoltre, sono emerse delle buone correlazioni tra tutte le dimensioni, fatta eccezione per la comprensibilità. Quest'ultimo elemento secondo le autrici è dato da un problema di interpretazione di tale descrittore. Nonostante i risultati positivi, è emerso uno scarso accordo assoluto tra i valutatori, che secondo le autrici è determinato da poche sessioni di *training*.

Infine, c'è uno studio di Pagliara (2017) nel quale è stata verificata l'applicabilità della scala su testi narrativi in italiano L2; gli informanti, i valutatori e l'organizzazione del *training* sono gli stessi dello studio di Faone e Pagliara (2017). I risultati dello studio hanno nuovamente confermato l'affidabilità della scala, sebbene anche in questo caso sia emerso uno scarso accordo assoluto tra i giudizi dei valutatori.

3. *Metodologia dello studio*

3.1. Obiettivi e domande di ricerca

Il presente studio, partendo dai suggerimenti proposti da Kuiken e Vedder (2017) e dai risultati degli studi che hanno avuto come lingua oggetto l'italiano (si veda § 2.1), si è posto l'obiettivo di testare l'applicabilità della scala non più su un singolo tipo di testo, ma su tre tipi di testi differenti: narrativi, regolativi, argomentativi.

Per raggiungere quest'obiettivo sono state formulate tre domande di ricerca:

1. C'è *interrater reliability* e *interrater agreement* tra i giudizi dei valutatori?
2. C'è correlazione tra i giudizi dei valutatori sulle quattro dimensioni di adeguatezza funzionale?
3. C'è correlazione tra i giudizi dei valutatori nei tre testi scritti da uno stesso informante?

Con la prima domanda ci si occupa di verificare l'affidabilità della scala come strumento di valutazione. La seconda domanda, invece, mira a verificare la possibilità di valutare la FA in termini di quattro dimensioni. Infine, la terza domanda, mira a verificare se la scala è applicabile in egual modo a tutti e tre i tipi di testi o se è influenzata dal tipo di testo valutato. Quest'ultimo elemento ha rappresentato la vera innovazione dello studio, in quanto per la prima volta è stato possibile confrontare le valutazioni date dagli stessi valutatori a tre testi prodotti dagli stessi informanti per ottenere informazioni circa l'applicabilità della scala su diversi tipi di testi. Di seguito sarà presentata la metodologia dello studio in riferimento alle modalità di raccolta (§ 3.2; § 3.3) e analisi dei dati (§ 3.4). Infine saranno presentati e discussi i risultati dell'analisi dei dati per poter rispondere alle domande di ricerca (§ 4; § 5).

3.2. Gli informanti e la raccolta dati

Per condurre il presente studio sono stati raccolti i testi scritti da 15 parlanti non nativi di italiano che al momento della raccolta dati erano iscritti al primo anno di *Italian Studies* presso l'Università di Amsterdam. Gli informanti erano parlan-

ti nativi di olandese – fatta eccezione per una madrelingua inglese e una bilingue olandese-francese – con un'età compresa tra i 19 e i 55 anni (età media 27,4). Gli informanti avevano un livello di italiano compreso tra A2 a B2, ma per avere una descrizione più precisa del loro livello di competenza linguistica si è deciso di far svolgere loro un C-test (Babaii - Ansari, 2001). Il C-test – fornito dall'Università di Amsterdam – è consistito nel completamento di 100 parole contenute in 5 testi brevi. Per poter completare le parole gli informanti hanno potuto affidarsi esclusivamente agli indizi contestuali senza poter utilizzare il dizionario. Per definire globalmente i livelli di competenza degli informanti ci si è affidati alla suddivisione fatta da Kuiken *et al.* (2010): livello basso (0-40); livello intermedio (41-60); livello avanzato (61-100). Gli informanti del presente studio sono risultati tutti appartenenti a una fascia intermedia o avanzata di italiano, con punteggi superiori a 48.

Per elicitare i tre tipi di testi, invece, sono stati utilizzati tre *task* differenti. Un *task* narrativo che consisteva nella produzione di un breve testo in cui raccontare un particolare evento accaduto durante un viaggio di studio. Un *instruction task*, utilizzato per elicitare il testo regolativo, che consisteva nello scrivere un biglietto d'istruzioni da lasciare a una coppia che aveva preso in affitto una casa per una vacanza. Un *decision-making task* che è stato utilizzato per elicitare un testo argomentativo in cui si doveva scrivere un'e-mail al direttore degli studenti internazionali argomentando la scelta di uno tra tre alloggi disponibili per un soggiorno di studio all'estero. Ogni *task* aveva una breve consegna e una lista di elementi da dover scrivere nel testo. Inoltre, per ogni *task* bisognava produrre un testo di almeno 150 parole senza la possibilità di utilizzare il dizionario. I *task* utilizzati in questo studio sono stati adattati in italiano a partire da tre task in lingua inglese ideati da Sofía Martín-Laguna, post-dottoranda dell'Università Jaume I, ricercatrice ospite presso l'Università di Amsterdam. In totale ogni informante ha prodotto 3 testi – un testo narrativo, un testo regolativo, un testo argomentativo – per un totale di 45 testi raccolti.

La raccolta dati ha avuto luogo in una singola sessione, corrispondente a una lezione di un corso di storia della lingua italiana. Gli informanti hanno avuto in totale un'ora e mezza per svolgere, nell'ordine: un breve questionario biografico-linguistico, i tre *task*, il C-test. Non è stato dato un limite di tempo per ogni attività, ma un tempo complessivo di svolgimento della prova, così da permettere agli studenti di gestire il lavoro in autonomia. In un primo momento è stato consegnato il plico con il questionario e i *task*; una volta conclusa questa parte gli informanti hanno ricevuto il C-test. Si è deciso di sequenziare le attività in questo modo per far sì che gli informanti svolgessero prima l'attività con il maggior carico cognitivo e alla fine della sessione quella con il minore.

3.3. I valutatori e la valutazione

Per poter testare la scala i testi raccolti sono stati fatti valutare da 5 valutatori non esperti (V1-V5), parlanti nativi di italiano, iscritti a un corso di laurea triennale o magistrale in lingue presso l'Università degli Studi Roma Tre. La valutazione dei testi è stata articolata in tre fasi: *training*, valutazione, post-valutazione. Per la valutazione dei

testi è stata utilizzata una versione tradotta in italiano della scala di Kuiken e Vedder (2017).

Il *training* ha avuto una durata totale di un mese circa ed è stato suddiviso in un *training* di gruppo e un *training* individuale. Il *training* di gruppo si è tenuto in presenza ed è iniziato con la presentazione delle scale. Una volta verificato che i valutatori avessero compreso i descrittori, sono stati fatti valutare loro quattro testi – tre di parlanti non nativi di italiano e uno di parlante nativo – per far sì che prendessero familiarità con le scale. Invece, nella sessione di *training* individuale, che si è tenuta online, i valutatori hanno dovuto valutare ulteriori dieci testi – cinque di parlanti nativi di italiano e cinque di non nativi – motivando i giudizi dati a ogni dimensione di ogni testo. Sia per il *training* online che per quello in presenza sono stati utilizzati testi argomentativi. I testi dei parlanti non nativi erano stati prodotti da madrelingua olandesi e gentilmente concessi per questa ricerca dall'Università di Amsterdam. I testi dei parlanti nativi, invece, appartenevano al corpus dello studio di Cortés Velásquez e Nuzzo (2017) ed erano stati prodotti da studenti dell'Università degli Studi Internazionali di Roma.

La fase di valutazione dei testi raccolti per lo studio, invece, è durata poco più di un mese e si è tenuta esclusivamente online. I testi sono stati suddivisi per tipologia e inviati ai valutatori in tre moduli a cadenza settimanale. Dall'invio dell'ultimo modulo alla fine della valutazione i valutatori hanno avuto altre tre settimane per completare il lavoro. In questa fase i valutatori non hanno dovuto dare una motivazione esplicita dei voti dati, ma semplicemente appuntarsi eventuali dubbi o riflessioni sul lavoro svolto.

Infine, la fase di post-valutazione è consistita in un'intervista retrospettiva in presenza durante la quale i valutatori hanno potuto esprimere le loro impressioni sul lavoro svolto e grazie alla quale sono emersi i punti di forza e le eventuali criticità della scala.

3.4. L'analisi dei dati

Per poter rispondere alle domande di ricerca i giudizi dati dai valutatori applicando la scala sono stati sottoposti a trattamenti statistici utilizzando il programma SPSS.

Per rispondere alla prima domanda sono state fatte due analisi statistiche: l'*interrater reliability* e l'*interrater agreement*. L'*interrater reliability* ha permesso di verificare la presenza di *consistency* tra i giudizi dei valutatori, ossia se questi concordassero tra loro nell'ordinare gerarchicamente i soggetti valutati (LeBreton - Senter, 2008: 816). L'*interrater agreement*, invece, ha permesso di verificare se i punteggi forniti dai valutatori fossero equivalenti in termini di valore assoluto del punteggio espresso (LeBreton - Senter, 2008: 816). L'*interrater reliability* è stata calcolata con l'indice alpha di Cronbach; *l'interrater agreement,* invece, è stato calcolato con i coefficienti di correlazione intraclasse. In entrambi i casi sono stati considerati alti valori al di sopra dello 0,7. Per ogni testo l'alpha di Cronbach e i coefficienti di correlazione intraclasse sono stati calcolati sui singoli punteggi dati alle quattro dimensioni di adeguatezza funzionale di ogni testo.

Per rispondere alla seconda e alla terza domanda è stata calcolata la correlazione tra le quattro dimensioni (domanda 2) e tra i giudizi dei valutatori nei tre testi scritti da

uno stesso informante (domanda 3). Con il termine correlazione si intende il processo statistico che misura il livello in cui i valori di due o più variabili sono connessi e come i cambiamenti nelle due variabili sono collegati tra loro (Abbott - McKinney, 2013: 127). Nel presente studio la correlazione è stata calcolata con il coefficiente di correlazione di Pearson e anche per questo indice statistico sono state considerate forti le correlazioni al di sopra dello 0,7. Inoltre, per determinare la significatività della correlazione ci si è affidati al *p value*, indicando come significative tutte le correlazioni con un *p value* inferiore a 0,05; al contrario, tutte le correlazioni con un *p value* superiore a 0,05 sono state considerate non significative. Per rispondere alla seconda domanda il coefficiente di correlazione di Pearson è stato calcolato sulle medie dei giudizi dati a ogni dimensione di ogni testo. Per rispondere alla terza domanda, invece, il coefficiente di correlazione di Pearson è stato calcolato sul giudizio medio di adeguatezza funzionale di ogni valutatore per ogni testo.

Inoltre, per poter verificare ulteriormente la possibilità di applicare la scala su tutti e tre i tipi di testi, i risultati delle analisi statistiche sono stati confrontati tra di loro.

Infine, per riuscire a interpretare meglio i risultati dell'analisi statistica, sono state analizzate qualitativamente le riflessioni fatte dai valutatori durante l'intervista retrospettiva dopo la valutazione.

4. *Risultati*

4.1. *Interrater reliability* e *interrater agreement*

L'analisi dell'*interrater reliability* e dell'*interrater agreement* ha riportato risultati contrastanti. Per quanto riguarda l'*interrater reliability*, dall'analisi dei dati sono emersi punteggi alti in tutti e tre i testi, dimostrando la presenza di una buona *consistency* tra i giudizi dei valutatori (tab. 1).

Tabella 1 - *Risultati del calcolo dell'alpha di Cronbach nei tre testi*

	Narrativo	*Regolativo*	*Argomentativo*
Contenuto (Cont.)	,894	,867	,861
Requisiti del tipo di task (Req.)	,876	,886	,914
Comprensibilità (Comp.)	,952	,934	,876
Coerenza e coesione (Coer./Coes.)	,870	,728	,774

Il testo narrativo ha ottenuto punteggi con valori che vanno da ,870 (coerenza e coesione) a ,952 (comprensibilità). Il testo regolativo, invece, ha ottenuto valori che vanno da ,728 (coerenza e coesione) a ,934 (comprensibilità). Nel testo argomentativo, infine, i valori dell'alpha di Cronbach nelle singole dimensioni vanno da ,774 (coerenza e coesione) a ,914 (requisiti del tipo di *task*).

Contrariamente a quanto emerso per l'*interrater reliability,* l'analisi dell'*interrater agreement* ha ottenuto valori abbastanza bassi in tutti e tre i testi (tab. 2).

Tabella 2 - *Risultati del calcolo dei coefficienti di correlazione intraclasse nei tre testi*

	Narrativo	*Regolativo*	*Argomentativo*
Cont.	,579	,400	,320
Req.	,540	,517	,548
Comp.	,786	,663	,565
Coer./Coes.	,496	,297	,365

Soltanto nel testo narrativo, infatti, i coefficienti di correlazione intraclasse hanno registrato un punteggio sopra la soglia dello 0,7 in una singola dimensione (comprensibilità: ,786). Al contrario, tali coefficienti sono risultati bassi nelle altre dimensioni, con valori che vanno da ,496 (coerenza e coesione) a ,579 (contenuto).

Il testo regolativo, invece, non ha mai superato la soglia dello 0,7 registrando valori bassi nei requisiti del tipo di *task* (,517) e nella comprensibilità (,663) e particolarmente bassi nel contenuto (,400) e nella coerenza e coesione (,297).

Nemmeno il testo argomentativo ha superato la soglia dello 0,7 registrando valori bassi nei requisiti del tipo di *task* (,548) e nella comprensibilità (,565) e particolarmente bassi nel contenuto (,320) e nella coerenza e coesione (,365).

4.2. Correlazione tra le quattro dimensioni

L'analisi del coefficiente di correlazione di Pearson nei tre testi ha riportato risultati interessanti e quasi sempre significativi (tab. 3).

Tabella 3 - *Risultati del calcolo del coefficiente di correlazione di Pearson tra le quattro dimensioni nei tre testi*

Narrativo			
	Req.	Comp.	Coer./Coes.
Cont.	,856**	,640*	,793**
Req.		,650**	,815**
Comp.			,791**
Regolativo			
	Req.	Comp.	Coer./Coes.
Cont.	,870**	,552*	,612*
Req.		,641*	,781**
Comp.			,781**
Argomentativo			
	Req.	Comp.	Coer./Coes.
Cont.	,772**	,499	,732**
Req.		,152	,695**
Comp.			,459

**p<0.01 *p<0.05

Nel testo narrativo tutte le correlazioni tra le quattro dimensioni hanno ottenuto punteggi significativi (p<0,05). È risultata particolarmente forte la correlazione tra contenuto e requisiti del tipo di *task* (,856). Altrettanto forte è risultata la correlazione tra coerenza e coesione e le altre dimensioni. È risultata più debole, ma altrettanto buona, la correlazione tra la comprensibilità con le dimensioni del contenuto (,640) e dei requisiti del tipo di *task* (,650).

Anche la correlazione tra le dimensioni nel testo regolativo è risultata significativa in tutti i casi (p<0,05). È risultata particolarmente forte la correlazione tra contenuto e requisiti del tipo di *task* (,870). Altrettanto forte è risultata la correlazione tra coerenza e coesione e requisiti del tipo di *task* (,781) e comprensibilità (,781). È risultata essere più debole, invece, la correlazione tra coerenza e coesione e contenuto (,612) e tra la comprensibilità e le dimensioni del contenuto (,552) e dei requisiti del tipo di *task* (,641).

A differenza di quanto si verifica per i due testi appena esaminati, nel testo argomentativo la correlazione tra le quattro dimensioni non sempre è risultata significativa. Infatti, sono risultate forti e significative (p<0.05) solo le correlazioni tra contenuto e requisiti del tipo di *task* (,772) e tra la coerenza e coesione con le dimensioni del contenuto (,732) e dei requisiti del tipo di *task* (,695). Al contrario, le correlazioni tra la comprensibilità e le altre dimensioni non sono mai risultate significative, con valori nettamente al di sotto della soglia dello 0,7 (,499 con il contenuto; ,152 con i requisiti; ,459 con la comprensibilità).

4.3. Correlazione tra i giudizi dei valutatori

Il calcolo del coefficiente di correlazione di Pearson tra i giudizi dei valutatori (V1-V5) nei tre testi ha riportato punteggi significativi in tutti i casi (tab. 4).

Tabella 4 - *Risultati del calcolo del coefficiente di correlazione di Pearson tra i giudizi dei valutatori nei tre testi*

Narrativo				
	V2	V3	V4	V5
V1	,883**	,837**	,844**	,826**
V2		,743**	,780**	,819**
V3			,802**	,883**
V4				,847**
Regolativo				
	V2	V3	V4	V5
V1	,832**	,585*	,679**	,636*
V2		,815**	,736**	,861**
V3			,704**	,829**
V4			,687**	
Argomentativo				
	V2	V3	V4	V5
V1	,784**	,611*	,779**	,905**
V2		,695**	,691**	,723**
V3			,579*	,549*
V4				,870**

**p<0.01 *p<0.05

Per quanto riguarda il testo narrativo, è risultata particolarmente forte la correlazione tra V1 e gli altri valutatori e tra V5 e gli altri valutatori. Altrettanto forte è risultata la correlazione tra V4 e V3. Più debole, ma comunque significativa, è risultata la correlazione tra V2 e V3 e V2 e V4.

La correlazione tra i giudizi dei valutatori nel testo regolativo, invece, è risultata particolarmente alta tra V2 e gli altri valutatori e tra V3 e gli altri valutatori. È risultata essere più debole, invece, la correlazione tra V1 e gli altri valutatori e tra V5 e V4.

Infine, la correlazione tra i giudizi dei valutatori nel testo argomentativo è risultata particolarmente forte tra V1 e gli altri valutatori. Altrettanto forte è risultata la correlazione tra V5 e le gli altri valutatori e tra V2 e gli altri valutatori. È risultata debole, invece, la correlazione tra V3 e gli altri valutatori e tra V4 e V2.

5. *Discussione dei risultati*

L'analisi statistica dei dati ha permesso di rispondere alle domande di ricerca e di individuare implicazioni interessanti circa l'applicabilità della scala su diversi tipi di testi.

Per quanto riguarda la prima domanda, si è potuto constatare che la scala è uno strumento abbastanza affidabile. Infatti, il calcolo dell'alpha di Cronbach relativo all'*interrater reliability* ha ottenuto dei valori alti in tutti e tre i testi e in tutte le dimensioni. Tra le quattro dimensioni quelle ad avere un valore di *interrater reliability* più alto sono i requisiti del tipo di *task* (testo argomentativo) e la comprensibilità (testo narrativo e regolativo). Nonostante questo risultato positivo, però, è emerso un aspetto che solleva dubbi circa l'affidabilità della scala: i bassi punteggi di *interrater agreement* registrati nei tre tipi di testi. Le dimensioni ad avere un *interrater agreement* più bassa sono il contenuto e la coerenza e coesione, che hanno ottenuto punteggi molto inferiori alla soglia dello 0,7. Questo risultato è in linea con quanto emerso dall'intervista retrospettiva con i valutatori, i quali hanno dichiarato di aver avuto molte difficoltà nell'interpretare e applicare i descrittori relativi a queste due dimensioni. Inoltre, questi risultati sono anche in linea con quelli emersi in Faone e Pagliara (2017) e Pagliara (2017): in entrambi gli studi erano stati registrati dei valori abbastanza bassi di *interrater agreement*. La mancanza di accordo assoluto potrebbe attribuirsi generalmente a problemi di interpretazione e/o formulazione dei descrittori o alla necessità di aumentare le sessioni di *training* per far sì che i valutatori prendano più familiarità con le scale e raggiungano un maggiore accordo.

I risultati dell'analisi dei dati hanno permesso di rispondere in maniera positiva anche alla seconda domanda di ricerca, confermando la possibilità di valutare la FA in termini di quattro dimensioni. È risultata essere particolarmente forte e significativa in tutti e tre i testi la correlazione tra contenuto e requisiti del tipo di *task*. Altrettanto forte è risultata essere la correlazione tra coerenza e coesione e le altre dimensioni. In riferimento a questo ultimo dato, rappresenta un'eccezione la correlazione tra coerenza e coesione e contenuto nel testo regolativo, che è leggermente

più debole rispetto agli altri due testi. Inoltre, è risultata particolarmente debole la correlazione tra comprensibilità e contenuto e comprensibilità e requisiti del tipo di *task*. Queste correlazioni deboli sono plausibili e non indicano un problema con la dimensione della comprensibilità, in quanto un testo comprensibile può non avere contenuti adeguati e non rispettare i criteri di un *task*, e viceversa. Inoltre, i valutatori hanno dichiarato di non aver avuto alcun problema nell'interpretazione dei descrittori di tale dimensione. I dati fin qui emersi sono parzialmente in accordo con quelli emersi in Faone e Pagliara (2017) e in Pagliara (2017).

Non tutti e tre i testi esaminati nel presente studio, però, hanno registrato correlazioni forti e significative. Il testo argomentativo, infatti, ha riportato delle correlazioni non significative, non solo nelle dimensioni che avevano una correlazione debole anche negli altri testi (comprensibilità/contenuto; comprensibilità/requisiti del tipo di *task*), ma anche nella correlazione tra comprensibilità e coerenza e coesione, che negli altri due testi era molto forte. Questo dato risulta particolarmente problematico, in quanto questi due aspetti sono strettamente collegati tra loro: un testo poco coerente e coeso difficilmente può essere facilmente comprensibile. Nonostante questo aspetto negativo, però, si è risposto positivamente alla seconda domanda, in quanto essendo emerse delle correlazioni non significative soltanto nel testo argomentativo, queste sono state collegate più a problemi di applicazione della scala su questo tipo di testo che a problemi relativi alla possibilità di valutare il costrutto in termini di quattro dimensioni. Inoltre, bisogna sottolineare che questi dati negativi sono in contrasto con quanto emerso nello studio di Kuiken e Vedder (2017), nel quale tutte le correlazioni raggiungevano la significatività.

È stato possibile rispondere positivamente anche alla terza domanda, in quanto le correlazioni tra i giudizi dei valutatori nei tre testi sono risultate essere sempre forti e significative. Ciò nonostante, confrontando tali correlazioni tra di loro è emerso che queste variano da un testo all'altro. In altre parole, le correlazioni più o meno forti tra coppie di valutatori sono risultate variabili al variare della tipologia testuale. Questo risultato ha portato a dedurre che le scale potrebbero essere influenzabili dal tipo di testo valutato e quindi non applicabili ugualmente a tutte e tre le tipologie. Quest'ultima deduzione è confermata anche da altri elementi, primo fra tutti i commenti fatti nell'intervista retrospettiva dai valutatori, i quali hanno affermato di aver avuto difficoltà ad applicare allo stesso modo i descrittori delle scale a tre tipi di testi diversi. Inoltre, se si confrontano complessivamente i risultati dell'analisi dell'affidabilità, questi sono diversi da un testo a un altro e questa diversità potrebbe essere data dall'impossibilità di applicare la scala a tipi di testi diversi. Infine, ad avvalorare ulteriormente l'ipotesi della sensibilità della scala al tipo di testo valutato sono le correlazioni non significative emerse nel testo argomentativo, ma non negli altri due tipi di testi.

Bisogna sottolineare che in questo studio è stata utilizzata una versione tradotta in italiano delle scale originali, pertanto i problemi emersi dall'analisi dei dati in riferimento alla formulazione dei descrittori e all'applicabilità delle scale a diversi tipi di testo potrebbero essere dati da problemi di traduzione. Un aspetto che, però, po-

trebbe essere problematico anche nella versione originale delle scale riguarda la coerenza e coesione; in quanto i valutatori stessi hanno ammesso di aver avuto difficoltà nel valutare la coerenza e la coesione facendo riferimento a un unico descrittore, poiché non sempre i due aspetti si trovavano sullo stesso livello, ritrovandosi spesso a dover sacrificare il punteggio dell'uno o dell'altro per attribuire un voto unico.

Complessivamente, quindi, l'analisi statistica dei dati ha permesso di rispondere positivamente a tutte le domande, confermando la possibilità di utilizzare la scala per valutare testi scritti, sebbene la sua formulazione attuale non sia applicabile ugualmente a tutti e tre i tipi di testi esaminati.

6. *Conclusioni*

Il presente studio ha avuto come obiettivo verificare l'applicabilità della scala di valutazione della FA ideata da Kuiken e Vedder (2017) su tre tipi di testi diversi (narrativi, regolativi, argomentativi). Per poter rispondere alle domande di ricerca sono state analizzate statisticamente le valutazioni date utilizzando la scala da parte di 5 valutatori non esperti parlanti nativi di italiano a 45 testi scritti da parlanti non nativi.

I risultati dell'analisi statistica hanno permesso di confermare che la FA può essere valutata in termini di quattro dimensioni, in quanto queste hanno riportato delle correlazioni forti e significative. Inoltre, la scala è risultata essere uno strumento abbastanza affidabile per valutare il costrutto.

Ciò nonostante, dall'analisi dei dati sono emersi anche alcuni aspetti critici delle scale, primo fra tutti lo scarso accordo assoluto tra i valutatori, che ha inciso negativamente sull'analisi dell'affidabilità dello strumento di valutazione. Inoltre, nel testo argomentativo sono emersi dei punteggi non significativi nelle correlazioni tra la comprensibilità e le altre dimensioni. A questi due aspetti negativi, va ad aggiungersi la variabilità delle correlazioni tra i giudizi dei valutatori nei tre testi e la generale variabilità dei risultati statistici da un testo all'altro.

Confrontando questi risultati negativi con le riflessioni fatte dai valutatori dopo lo svolgimento dell'attività, è stato dedotto che le criticità emerse potrebbero essere attribuite a problemi di formulazione dei descrittori delle scale nella versione italiana. Secondo quanto affermato dai valutatori, infatti, i descrittori risultano difficilmente comprensibili in riferimento ad alcuni aspetti di alcune dimensioni e non applicabili in egual modo a tutti i tipi di testi. Un ulteriore aspetto problematico potrebbe essere riferito alla modalità di svolgimento del *training*, probabilmente non sufficiente per far sì che i valutatori prendessero familiarità con l'utilizzo delle scale.

Alla luce dei risultati presentati, pertanto, sarebbe opportuno che in studi futuri, prima di testare nuovamente l'applicabilità delle scale, si apportassero alcune modifiche per cercare di ovviare ai problemi emersi dall'analisi statistica. Innanzitutto, sarebbe opportuno apportare alcune modifiche alla formulazione delle scale, soprattutto nelle dimensioni che sono risultate più problematiche sia in base ai dati statistici che in base a quanto hanno affermato i valutatori durante l'intervista retro-

spettiva. Per quanto riguarda il contenuto e i requisiti del tipo di *task*, trattandosi di problemi di formulazione dei descrittori, bisognerebbe trovare delle formulazioni che rendano la scala idonea ad essere usata su diversi tipi di testi. Per quanto riguarda, invece, la coerenza e coesione, dato che i valutatori hanno avuto difficoltà a valutare i due aspetti affidandosi ad una singola scala, sarebbe opportuno provare a dividere la dimensione in due sotto-dimensioni, cercando di valutare i due aspetti in maniera separata.

Inoltre, sarebbe opportuno rivedere anche la metodologia di svolgimento del *training*. L'aver tenuto un'unica sessione in presenza e aver utilizzato un solo tipo di testo può non essere stato sufficiente a far sì che i valutatori raggiungessero un buon accordo assoluto. Di conseguenza, in studi futuri si potrebbe provare a risolvere il problema aumentando le sessioni di *training*, utilizzando tipi di testi diversi, come in effetti raccomandano Kuiken e Vedder (2017).

Infine, visti i dati negativi relativi al testo argomentativo e dato che questi sono in contrasto con quanto emerso in Kuiken e Vedder (2017), in studi futuri sarebbe opportuno fare un'analisi più approfondita su questo tipo di testo, testando eventualmente la scala su diversi tipi di testi argomentativi, per verificare se si tratta di un problema di scala o di un problema di tipologia testuale.

I risultati di questo studio sono emersi dall'analisi di un numero ristretto di dati, pertanto non possono essere generalizzabili. Ciò nonostante, si spera che quanto emerso possa avere delle implicazioni positive sull'applicabilità della scala e fornire spunti per studi futuri.

6. *Ringraziamenti*

Il presente lavoro è frutto di una collaborazione tra l'Università di Roma Tre e l'Università di Amsterdam ed è stato possibile grazie ad una borsa di ricerca per tesi erogata dall'Università di Roma Tre. Si ringraziano la professoressa Elena Nuzzo e la professoressa Ineke Vedder che hanno avuto il ruolo di supervisori nella raccolta e nell'analisi dei dati. Infine, si ringrazia il professor Giuseppe Bove per il supporto relativo all'analisi statistica dei dati.

Bibliografia

Abbott M.L. - McKinney J. (2013), *Understanding and applying research design*, John Wiley & Sons, Hoboken.

Babaii E. - Ansari H. (2001), The C-Test: a valid operationalization of reduced redundancy principle?, in *System*, 29(2): 209-219.

Bridgeman B. - Powers D. - Stone E. - Mollaun P. (2011), TOEFL iBT speaking test scores as indicators of oral communicative language proficiency, in *Language Testing*, 29(1): 91-108.

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

Cortés Velásquez D. - Nuzzo E. (2017), *Assessing task performance in the academic context: functional adequacy scales for Italian L1 university students*, 7th TBLT Conference Tasks in Context, University of Barcelona, Barcelona, April 19-21.

Council of Europe (2018), *Common European Framework of Reference for Languages: learning, teaching, assessment – Companion volume with new descriptors,* https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

De Jong N.H. - Steinel M.P. - Florijn A.F. - Schoonen R. - Hulstijn, J.H. (2012a), Facets of speaking proficiency, in *Studies in Second Language Acquisition*, 34(1): 5–34.

De Jong N.H. - Steinel M.P. - Florijn A.F. - Schoonen R. - Hulstijn, J.H. (2012b), The effect of task complexity on functional adequacy, fluency and lexical diversity in speaking performances of native and non-native speakers, in Housen A. - Kuiken F. - Vedder I. (eds), *Dimensions of L2 performance and proficiency: complexity, accuracy and fluency in SLA*, John Benjamins, Amsterdam: 121–142.

Faone S. - Pagliara F. (2017), *How to assess L2 information-gap tasks through functional adequacy rating scales*, 7th TBLT Conference Tasks in Context, University of Barcelona, Barcelona, April 19-21.

Grice H.P. (1975), Logic and conversation, in Cole P. - Morgan, J.L .(eds), *Syntax and Semantics 3: Speech acts*, Academic Press: New York: 41-58.

Hismanoglu M. (2011), An investigation of ELT students' intercultural communicative competence in relation to linguistic proficiency, overseas experience and formal instruction, in *International Journal of Intercultural Relations*, 35(6): 805–817.

Kuiken F. - Vedder I. (2014), Rating written performance: what do raters do and why?, in *Language Testing*, 31(3): 329-348.

Kuiken F. - Vedder I. (2017), Functional adequacy in L2 writing: towards a new rating scale, in *Language Testing*, 34(3): 321-336.

Kuiken F. - Vedder I. - Gilabert R. (2010), Communicative adequacy and linguistic complexity in L2 writing, in Bartning I. - Martin M. - Vedder I. (eds), *Communicative proficiency and linguistic development: intersections between SLA and language testing research*, Eurosla Monographs Series 1: 81-100.

LeBreton J.M. - Senter J.L. (2008), Answers to 20 questions about interrater reliability and interrater agreement, in *Organizational Research Methods*, 11 (4): 815-852.

Pagliara F. (2017), *Valutare l'adeguatezza funzionale in produzioni scritte di studenti Marco Polo*, Convegno "Dieci anni di didattica dell'italiano a studenti cinesi. Risultati, esperimenti, proposte", Università per Stranieri di Siena, Siena, 6-7 ottobre.

Pallotti G. (2009), CAF: Defining, refining and differentiating constructs, in *Applied Linguistics*, 30(4): 590-601.

Vedder I. (2016), Il ruolo dell'adeguatezza funzionale nelle produzioni scritte in lingua seconda: proposta per una scala di valutazione, in Vedder I. - Santoro E. (a cura di), *Pragmatica e interculturalità in italiano lingua seconda*, Franco Cesati, Firenze: 79-92.