MISURARE L'ADEGUATEZZA FUNZIONALE IN TESTI SCRITTI DI APPRENDENTI DI ITALIANO L2

Paolo Orrù[1]

1. INTRODUZIONE

L'interesse per la dimensione funzionale del linguaggio e per gli aspetti pragmatici coinvolti nell'apprendimento di una L2 ha iniziato a trovare in anni recenti una proficua sistematizzazione[2]. Nonostante le capacità comunicative e la padronanza della lingua in uso siano da anni al centro della proposta del *Quadro Comune Europeo di Riferimento per le Lingue*[3] (QCER), gli studi relativi alla misurazione e alla valutazione delle produzioni degli apprendenti si sono appoggiati preferenzialmente agli indici CAF (complessità, accuratezza, fluenza)[4], anche, e forse soprattutto, in assenza di una determinazione univoca di cosa sia la competenza comunicativa (vd. Sez. 1). Tali indici, pur essendo strumenti utilissimi per definire e testare le *performance* linguistiche, da soli non appaiono sufficienti a verificare l'adeguatezza di un testo ai suoi scopi comunicativi: «Receiving a high rating on a rating scale does not provide direct evidence that a particular message was understood» (Bridgeman *et al.*, 2012: 92).

Il presente studio intende, dunque, indagare alcuni aspetti della valutazione delle competenze comunicative di apprendenti di italiano L2 attraverso l'applicazione di una scala di valutazione delle produzioni scritte elaborata da Kuiken e Vedder[5] e messa alla prova in studi recenti in vari contesti e tipologie di apprendenti[6]. Si farà, dunque, riferimento al concetto di *adeguatezza funzionale*, inteso come un costrutto multidimensionale che intende incorporare l'efficacia della trasmissione del messaggio e l'esecuzione riuscita del compito.

La prima sezione illustrerà sinteticamente cosa si intenda con adeguatezza funzionale e quali siano le dimensioni che compongono il costrutto; la seconda sezione presenta la descrizione della scala e i principi che la orientano; la terza sezione passerà in rassegna le metodologie di raccolta del *corpus*, la sua composizione e le fasi di addestramento dei

[1] University of Debrecen (HU).

[2] Si veda a questo proposito la dettagliata rassegna di studi proposta da Nuzzo e Santoro (2017).

[3] E del suo recente aggiornamento Council of Europe (2018), che stabilisce ancora una volta la centralità di un'educazione linguistica basata su compiti reali.

[4] Come ricordato da Pallotti (2009: 596): «It is surprising how few CAF studies report data about the communicative success and adequacy of the tasks and the learners investigated».

[5] La scala è stata proposta nella sua versione italiana in Vedder (2016) e in inglese in Kuiken, Vedder (2017).

[6] Vedder (2016) e Kuiken, Vedder (2017) presentano i risultati dei primi esperimenti svolti con apprendenti di italiano e olandese L2 e con un campione di parlanti italiani e olandesi di madrelingua; Faone, Pagliara (in stampa); Pagliara (in stampa) con parlanti sinofoni. Altri esperimenti sono tutt'ora in svolgimento con parlanti di madrelingua spagnola, olandese e italiana. Kuiken, Vedder (2018) offre invece un confronto scritto-parlato per italiano e olandese L2 frutto di un esperimento pilota in cui le scale sono state adattate per dei task monologici.

valutatori e il processo di valutazione; la quarta sezione presenterà i risultati e la loro analisi statistica. Infine, si cercherà di trarre alcune conclusioni.

2. L'ADEGUATEZZA FUNZIONALE: UN COSTRUTTO MULTIDIMENSIONALE

Contrariamente alla misurazione degli indici di complessità sintattica, accuratezza, fluenza e varietà lessicale generalmente impiegati in L2 per vagliare le capacità degli apprendenti, non esistono allo stato attuale indici globali per misurare l'adeguatezza comunicativa[7]. Ciò è sicuramente in parte dovuto anche all'assenza di una definizione univoca di cosa debba intendersi con *adeguatezza comunicativa*, basti considerare l'abbondanza di etichette assegnate, seppur con sfumature e presupposti teorici differenti, allo stesso costrutto: *communicative adequacy* (Kuiken, Vedder, Gilabert, 2010; Pallotti, 2009; Révész *et al.*, 2016), *communicative competence* (McNamara, Roever, 2007), *communicative effectiveness* (Bridgeman *et al.*, 2012; Sato, 2012), *intercultural competence* (Hismanoglu, 2011), *communicative functionality* (Fragai, 2001).

Per Bridgeman *et al.* (2012) essa va intesa come un indicatore della comprensibilità della produzione di un parlante da parte di un interlocutore, mentre per McNamara, Roever (2007) è vista in termini di appropriatezza socio-pragmatica della produzione linguistica al contesto specifico. Pallotti (2009), Kuiken *et al.* (2010) e De Jong *et al.* (2012a, 2012b) fanno, poi, diretto riferimento alla riuscita esecuzione di un compito comunicativo (in termini di aderenza del contenuto alle istruzioni di un task) e Upshur, Turner (1995) all'avvenuto trasferimento di informazioni. Knoch (2009), infine, considera il costrutto da un punto di vista discorsivo e lo definisce in termini di coerenza e coesione del testo.

Da questa sintetica disamina si può cogliere come le varie definizioni dell'adeguatezza (o efficacia) comunicativa non siano certamente del tutto sovrapponibili. Ognuna di queste etichette si focalizza su aspetti assai differenti della produzione in L2 e della sua valutazione: vengono chiamati in causa una pluralità di aspetti (socio)linguistici, pragmatici e testuali da tenere inevitabilmente in considerazione nell'elaborazione di indici oggettivi per la valutazione.

Partendo da queste riflessioni, Vedder (2016: 80) fornisce la seguente definizione:

> [un] costrutto interazionale, inteso in termini di 'felice' esecuzione del task e ispirato alle massime conversazionali di Grice. Vista in questa ottica, l'adeguatezza funzionale è determinata dal successo della trasmissione del messaggio del parlante A recepito dall'interlocutore B.

Impossibile non cogliere la natura eminentemente pragmatica del costrutto così definito, non solo per via dei riferimenti alle matrici *searliane* e *griciane*, ma anche grazie alla centralità assegnata al contesto di enunciazione e all'avvenuta trasmissione del messaggio. Lo spostamento terminologico da "comunicativa" a "funzionale" è utile, invece, a metter ancor più in risalto la relazione diretta tra il felice completamento del task[8] e l'adeguatezza

[7] È chiaramente necessario prestare particolare cautela nella formulazione di una scala di valutazione, poiché la decisione di cosa includervi ha ovvie ripercussioni sul giudizio finale assegnato all'apprendente e alla validità del test in sé; è inoltre impossibile valutare in un singolo test tutti gli aspetti della performance e della competenza linguistico-comunicativa di un informante. La scala deve essere quindi pensata avendo in mente una tipologia di *testing* ben precisa (Sato, 2012).

[8] «As a descriptor of communicative success and efficiency, adequacy can be measured in several ways. In closed tasks with correct/incorrect outcomes, it can be rated straightforwardly as the ratio of correct items achieved. In open tasks with no predefined correct answer, adequacy can be evaluated by means of

della *performance* in L2, valutata in base allo specifico contesto linguistico del compito[9] (una lettera formale, un racconto, una lamentela, ecc.). L'obiettivo, pertanto, non è quello di descrivere una competenza globale dell'apprendente di una L2, quanto il descrivere la sua capacità di produrre testi adeguati alla specifica funzione comunicativa richiesta da un compito.

3. LE SCALE DI VALUTAZIONE

La scala di valutazione proposta da Kuiken e Vedder[10] ha le sue basi nella pragmatica, e più specificamente prende ampiamente spunto dalle massime conversazionali di Grice (1975): quantità, qualità, modo e relazione. L'adeguatezza funzionale, dunque, è vista come composta da quattro dimensioni: *contenuto*, *requisiti del task*, *comprensibilità*, *coerenza* e *coesione*. Come ben illustra Vedder (2016: 82): «[l]'assunto primario che sottostà alla scala è che non è possibile valutare la competenza scritta in L2 senza prendere in considerazione gli aspetti funzionali del testo». Come si vedrà, la scala cerca inoltre di inglobare i differenti costrutti teorici riconosciuti negli studi elencati in precedenza e l'impostazione suggerita dai descrittori delle scale del QCER (Council of Europe, 2001). Ogni dimensione è rappresentata da una scala Likert in sei punti con descrittori dettagliati, per la cui lettura integrale rimandiamo all'appendice.

Riportiamo di seguito le quattro dimensioni della scala e la loro descrizione sintetica[11]:

- *Contenuto*: questa dimensione prende in considerazione l'adeguatezza del numero e del tipo di unità informative espresse nel testo del parlante/scrivente A, la loro pertinenza e rilevanza, indipendentemente dai requisiti specifici riguardo all'esecuzione del task (Grice, 1975; massime di quantità, di relazione e di qualità).
- *Requisiti del task*: è stato dato questo nome al criterio che considera la misura in cui nel testo si è tenuto conto dei requisiti specifici del task che il parlante/scrivente A deve svolgere, relativi al genere testuale, all'atto linguistico, alla macro-struttura e al registro del testo (Grice, 1975; massime di relazione e di modo).
- *Comprensibilità*: qui si focalizza il grado di comprensibilità del testo di A e lo sforzo richiesto all'interlocutore B per capire lo scopo e le idee espresse nel testo (Grice, 1975; massima di modo).
- *Coerenza e coesione*: questo criterio si riferisce alla coerenza e alla coesione del testo del parlante/scrivente A, in termini della presenza o assenza di strategie anaforiche, connettivi, salti logici e argomenti non collegati, ripetizioni, ecc (Grice, 1975, massima di modo).

Benché formulate diversamente, le scale si ispirano all'impostazione del Quadro comune e alla logica dei livelli scalari dei suoi descrittori. In tale ottica, le scale per la competenza pragmatica, recentemente aggiornate dal *Companion volume* (Council of

qualitative ratings, using predefined descriptor scales like the ones of the Common European Framework of Reference» (Pallotti, 2009: 597).

[9] Cfr. Kuiken, Vedder (2017: 323).

[10] La scala è stata presentata ormai in più studi, si veda Vedder (2016), Kuiken, Vedder (2014, 2017 e 2018); sviluppata originariamente per indagare il rapporto tra complessità linguistica e adeguatezza funzionale in parlanti L2 e L1 e messa in pratica in un primo esperimento con parlanti di italiano, olandese e spagnolo L2 e L1 in Kuiken *et al.* (2010).

[11] Le descrizioni sono tratte da Vedder (2016: 82-83).

Europe, 2018: 136-ss), sono modellate sui livelli di competenza generali (da pre A1 a C2) e mirano a descrivere competenze globali del parlante[12], anche attraverso la segnalazione di specifici compiti comunicativi e atti linguistici; le scale elaborate da Kuiken e Vedder, invece, si focalizzano sull'adeguatezza della singola produzione comunicativa allo specifico compito comunicativo da eseguire. L'originalità della scala risiede proprio nella centralità assegnata all'aspetto funzionale dei testi; allo stesso tempo la misurazione avviene in maniera indipendente dai canonici indici CAF di complessità e accuratezza linguistica.

L'obiettivo di una scala così pensata e costruita è quella di offrire uno strumento che sia utilizzabile da valutatori esperti e non esperti; che risponda a misurazione il più possibile precise e non olistiche, per limitare il fattore soggettivo e impressionistico nella valutazione; che sia autonoma rispetto ai canonici criteri CAF; infine, che sia applicabile a testi di parlanti L1 o L2 indistintamente[13].

4. METODI, *CORPUS*, VALUTATORI

4.1. *Metodi e domande di ricerca*

Dal punto di vista del tema e delle domande di ricerca, l'esperimento è sostanzialmente una parziale replica delle prove precedenti; ciò si giustifica con la necessità di testare la scala per task di varia natura e scopo comunicativo, oltre a quelli argomentativi per cui è stata originariamente pensata e su cui è stata applicata inizialmente. Ulteriore elemento di interesse per la reiterazione del *testing* è la verifica della sua applicabilità per L2 e L1 differenti; nel nostro caso, assume maggior rilievo la possibilità di applicare lo strumento alle produzioni di parlanti di una madrelingua tipologicamente diversa rispetto alle ricerche già compiute o in corso di realizzazione.

Seguendo le precedenti esperienze a cui si è fatto riferimento sopra, abbiamo deciso di testare le scale di valutazione su un campione di studenti ungheresi di italiano LS (per la descrizione dettagliata del *corpus*, si veda 3.2). Nello specifico, il nostro obiettivo fondamentale è quello di esaminare l'affidabilità delle scale di valutazione di Kuiken e Vedder attraverso tre domande di ricerca:

1) C'è *interrater reliability* e *interrater agreement* tra i giudizi dei valutatori?
2) C'è correlazione tra le quattro dimensioni dell'adeguatezza funzionale?
3) C'è correlazione tra i livelli di competenza degli apprendenti e i giudizi dei valutatori nelle quattro dimensioni e nell'adeguatezza funzionale?

È stato necessario raccogliere un cospicuo numero di testi *input* da sottoporre al giudizio di 4 valutatori. Sono stati impiegati tre tipi di task scritti (rimandiamo alla sezione successiva per la descrizione dettagliata); ciò allo scopo evidente di verificare la validità dello strumento valutativo nel più ampio raggio possibile di testi. In seguito, è stato

[12] Ad esempio, flessibilità del parlante ad adattarsi alla situazione comunicativa; presa di turno; precisione della formulazione linguistica di quanto si voglia esprimere; sviluppo tematico; coerenza e coesione.
[13] Si vedano Kuiken, Vedder (2017, 2018) per una più ampia descrizione degli obiettivi alla base della formulazione delle scale.

somministrato un C-test[14] per registrare i livelli di competenza degli informanti: il test è composto da 5 testi contenente 100 parole da riempire, senza l'ausilio del dizionario.

4.2. *Corpus*

Il *corpus* è formato da 120 testi scritti basati su tre differenti task, ogni informante ha prodotto, dunque, altrettanti testi. I 40 informanti sono tutti studenti di madrelingua ungherese dei corsi di laurea triennale in Italianistica e di quello magistrale in Traduzione specialistica e letteraria del Dipartimento di Italianistica dell'Università di Debrecen.

Come si accennava sopra, gli informanti hanno dovuto svolgere tre task corrispondenti ad altrettante tipologie testuali: uno regolativo, uno argomentativo e uno narrativo. La lunghezza prevista per ogni testo era di almeno 150 parole e ognuno era corredato da una serie di requisiti a cui rispondere. Il primo consisteva nel fornire una serie di informazioni e istruzioni a una coppia di ospiti del proprio *bed and breakfast*; il secondo prevedeva la scrittura di una mail formale per motivare la propria scelta di un alloggio studentesco per un periodo di studi all'estero; il terzo, infine, richiedeva di raccontare, per un concorso di scrittura, una breve storia incentrata su un episodio avvenuto durante un viaggio di studio[15].

Per ogni gruppo di studenti (corrispondente a una classe) si è svolta una sessione di raccolta dei dati della durata di 90 minuti, al cui interno era prevista anche la compilazione del C-test. Nonostante gli studenti siano stati lasciati liberi di svolgere i compiti nell'ordine che più preferivano, le tracce sono state consegnate in sequenze differenziate, per eliminare possibili effetti esterni alla procedura.

Gli studenti del corso triennale hanno tutti studiato italiano già nella scuola secondaria per almeno 4 anni; il superamento di un esame di italiano durante la maturità è un requisito necessario per accedere al corso di laurea universitario in Italianistica. Tuttavia, gli informanti provengono da esperienze formative e personali piuttosto eterogenee. Una buona parte di loro ha frequentato un programma bilingue al cui interno l'insegnamento della lingua italiana è decisamente rilevante e in cui anche materie curricolari (come storia e geografia) vengono insegnate in italiano e in cui sono presenti anche docenti italiani. Gli altri studenti, invece, affrontano l'apprendimento come una seconda lingua straniera accanto all'inglese (soprattutto) o a un'altra lingua (principalmente tedesco), per circa 3/4 ore settimanali con insegnanti di madrelingua ungherese. Gli studenti di Laura magistrale (solamente 3 nel nostro *corpus*) hanno affrontato l'intero percorso di studi in Italianistica. I livelli di competenza dei nostri informanti sono, dunque, piuttosto vari e il nostro campione è stato scelto precisamente a tale scopo.

4.3. *Valutatori*

Per testare le scale si è deciso di selezionare 4 valutatori non esperti, tutti studenti di corsi di Laurea in lingue e letterature straniere dell'Università di Cagliari: essi possiedono

[14] Il C-test in questione è stato realizzato dall'Università di Amsterdam e impiegato anche negli esperimenti precedenti allo stesso scopo. Per una descrizione si veda Kuiken *et al.* (2010).

[15] I tre task sono stati sviluppati all'Università di Amsterdam e sono stati somministrati da Del Bono (2019) (con studenti olandesi di italiano L2) e in Spagna (con studenti di spagnolo L1 e inglese L2), i cui dati non sono stati ancora analizzati nel momento in cui si scrive.

quindi almeno una basilare formazione linguistica e padroneggiano i fondamentali concetti della pragmatica e della linguistica testuale, utili per meglio comprendere anche a una prima lettura i descrittori e l'impostazione degli strumenti. L'impiego di valutatori non esperti risponde a due obiettivi fondamentali: il primo è la comparabilità con altri studi fatti per testare le stesse scale; il secondo è l'esigenza di analizzare come le produzioni degli informanti vengano recepite e comprese da parlanti nativi considerabili pari agli informanti. Essi hanno decisamente meno dimestichezza con produzioni di apprendenti L2 rispetto agli esperti, e possono affrontare con meno aspettative l'intero processo di valutazione; studi precedenti hanno registrato come i valutatori esperti possano essere influenzati dalle proprie esperienze didattiche e di *testing*; ciò, tuttavia, non intende rappresentare in nessun modo una messa in discussione del ruolo dei valutatori esperti o del docente di lingua in sé[16].

Si è proceduto a una fase di addestramento e formazione all'impiego delle scale in due fasi. La prima è consistita in un incontro faccia a faccia con i valutatori in cui sono state lette le scale e spiegati nel dettaglio i descrittori, al fine ovviamente di illustrarne la *ratio* e chiarire possibili dubbi. Nello stesso incontro è stata effettuata una prova di valutazione somministrando 3 testi di madrelingua ungheresi, uno per ogni task, e altrettanti di parlanti nativi, come forma di controllo: la scala è potenzialmente applicabile anche a parlanti L1, e può non essere banale ricordare che un testo di un parlante nativo potrebbe risultare meno "adeguato" di quello di un parlante non nativo. Infine, sono state confrontate e discusse le valutazioni e si è risposto ad alcune domande e dubbi dei valutatori.

Per motivi puramente organizzativi, la seconda fase è stata invece affrontata individualmente ed è stata svolta online: a ogni valutatore sono stati assegnati ulteriori 6 testi (2 per ciascun task) di apprendenti e 3 (ancora, uno per task) di madrelingua. I valutatori hanno discusso e motivato le loro scelte con il formatore. I testi usati come prova sono stati elicitati nella fase di raccolta del *corpus* e tenuti fuori dall'analisi o sono stati prodotti da ex studenti del Dipartimento di Italianistica di Debrecen; i testi di madrelingua sono stati invece prodotti appositamente da studenti universitari italiani.

La fase di valutazione si è svolta interamente online, si è proceduto a inviare i testi input in tre scaglioni divisi per tipologia di testi. L'ordine dei testi input è stato reso casuale in sequenze *ad hoc* per ogni valutatore, così da ridurre qualsiasi effetto esterno dovuto alla distribuzione dei testi. I valutatori hanno dovuto riportare in un'apposita tabella i giudizi da 1 a 6, senza dover motivare in altro modo le loro scelte.

Infine, il processo si è concluso con un'intervista retrospettiva individuale, per cogliere meglio il modo in cui hanno lavorato con le scale; le motivazioni dietro ai giudizi espressi; eventuali problemi incontrati nella valutazione o suggerimenti.

5. RISULTATI

I dati così raccolti sono stati analizzati statisticamente via software (SPSS) per testare essenzialmente 4 aspetti: l'affidabilità generale della scala; la correlazione tra i giudizi dei valutatori; la correlazione tra le varie dimensioni del costrutto; la correlazione tra il livello di competenza e le misure dell'adeguatezza funzionale.

Per verificare i primi due aspetti si è deciso di calcolare due indici: l'*interrater reliability* e l'*interrater agreeement*. Il primo misura il grado di consenso registrato tra i punteggi assegnati

[16] Lo studio pilota effettuato con la prima versione della scala (Kuiken *et al.*, 2010) è stato realizzato proprio con l'ausilio di soli valutatori esperti.

ai singoli informanti da ciascun valutatore, viene usato quindi per verificare se i valutatori usino la scala in modo coerente tra loro; tale dato è stato ricavato attraverso il calcolo dell'alpha di Cronbach applicato alle singole dimensioni.

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

	Narrativo	**Regolativo**	**Argomentativo**
Contenuto	,836	,817	,852
Requisiti del task	,826	,787	,828
Comprensibilità	,863	,867	,834
Coerenza/coesione	,865	,880	,875

Tutte le misurazioni sono risultate estremamente significative (Tabella 1) da un punto di vista statistico ($p<,01$), vengono considerate accettabili le misure a un livello superiore a 0,7, buone quelle superiori a 0,8 ed eccellenti quelle sopra 0,9. I risultati sono tutto sommato omogenei, indipendentemente dal task e dalla dimensione, e globalmente buoni; spicca senza dubbio l'affidabilità per quanto riguarda la dimensione di coerenza/coesione.

L'*interrater agreement* determina, invece, il consenso assoluto tra i giudizi assegnati dai valutatori a ciascun informante, ovvero se il punteggio assegnato da un valutatore sia di fatto intercambiabile con quello degli altri. Ciò è stato valutato mediante il calcolo dei coefficienti di correlazione intraclasse per le singole dimensioni.

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

	Narrativo	**Regolativo**	**Argomentativo**
Contenuto	,792	,796	,835
Requisiti del task	,782	,773	,801
Comprensibilità	,833	,819	,794
Coerenza/coesione	,834	,792	,804

Tutte le correlazioni raggiungono risultati sufficienti (superiori a 0,7) o buoni (superiori a 0,8). Tra i tre compiti è il testo regolativo a riportare in media i valori più bassi, attestandosi comunque su risultati sufficienti; viceversa è il task argomentativo a mostrare complessivamente livelli di accordo migliori per 3 delle 4 dimensioni.

Per rispondere invece al secondo dei nostri quesiti (la correlazione tra le varie dimensioni del costrutto) si è proceduto al calcolo del coefficiente di correlazione di Pearson: sono stati messi a confronto i giudizi dei valutatori nelle singole dimensioni per riscontrare se ci sia di fatto un rapporto tra le 4 dimensioni.

In tutti e tre i task i valori registrati mostrano una significativa correlazione tra le varie dimensioni ($p<0,01$).

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

Narrativo			
	Requisiti del task	Comprensibilità	Coerenza/coesione
Contenuto	,939	,728	,887
Requisiti del task		,763	,878
Comprensibilità			,886
Regolativo			
	Requisiti del task	Comprensibilità	Coerenza/coesione
Contenuto	,933	,847	,875
Requisiti del task		,888	,902
Comprensibilità			,928
Argomentativo			
	Requisiti del task	Comprensibilità	Coerenza/coesione
Contenuto	,911	,873	,888
Requisiti del task		,890	,892
Comprensibilità			,971

Come si può evincere dalla Tabella 3, tutte le correlazioni raggiungono la significatività statistica. In generale è la relazione tra le prime due dimensioni (contenuto e requisiti del task) a dimostrarsi particolarmente forte, con valori globalmente molto buoni: da ,911 del task argomentativo a ,939 di quello narrativo. Se osserviamo nel dettaglio i singoli compiti, si può notare come i livelli più bassi, seppur ancora sufficienti, siano stati misurati per le correlazioni tra la comprensibilità e il contenuto (,728), da una parte, e i requisiti del task (,763) dall'altra.

Un altro valore globalmente elevato è quello del rapporto tra comprensibilità e coerenza/coesione in tutti i task, ma con misure molto buone soprattutto in quelli regolativi (,928) e argomentativi (,971). Tali risultati sono in linea con i rilievi di Vedder (2016), che hanno dimostrato una forte correlazione tra queste due dimensioni sia in italiano (,938) sia in olandese (,873).

Può essere interessante infine verificare se esista una correlazione significativa tra i livelli di competenza misurati attraverso il C-test (vd. 3.2) e l'adeguatezza funzionale nel suo insieme o le singole dimensioni. È bene premettere che i dati seguenti non possono essere sovrastimati o sovrainterpretati per vari motivi. Certamente non sorprenderà una correlazione tra padronanza linguistica e adeguatezza funzionale; allo stesso tempo il C-test, benché si sia rivelato uno strumento di misurazione della padronanza linguistica globale tutto sommato utile, all'opposto, esso sicuramente non è pensato per verificare la padronanza in termini di adeguatezza funzionale.

Tabella 4.1. *Correlazione di Pearson tra livello di competenza e adeguatezza funzionale*

	Narrativo	**Regolativo**	**Argomentativo**
C-test	,725	,745	,787

Come si può notare, la correlazione tra il valore del C-test e l'adeguatezza funzionale (misurata attraverso la media delle singole dimensioni) è sufficientemente importante per tutti e tre i task, soprattutto per quello argomentativo. Vediamo di seguito, invece, quale sia il rapporto con le singole dimensioni.

Tabella 4.2. *Correlazione di Pearson tra livello di competenza e dimensioni*

Narrativo				
	Contenuto	Requisiti del task	Comprensibilità	Coerenza/coesione
C-test	,601	,616	,739	,767
Regolativo				
	Contenuto	Requisiti del task	Comprensibilità	Coerenza/coesione
C-test	,646	,732	,723	,755
Argomentativo				
	Contenuto	Requisiti del task	Comprensibilità	Coerenza/coesione
C-test	,716	,743	,774	,797

Dalla Tabella 4.2 si evince come quasi tutte le misure siano sufficientemente significative. Il task argomentativo è l'unico tra i tre tipi di testi a ottenere risultati superiori a 0,7 in tutte e quattro le scale; gli indicatori, poi, sono più alti in ogni singola dimensione rispetto ai task regolativi e narrativi. Questi ultimi registrano, inoltre, valori non sufficienti per la dimensione del contenuto; infine, il task narrativo è l'unico a riportare una correlazione inferiore a 0,7 anche per i requisiti del task.

6. DISCUSSIONE E CONCLUSIONI

Si tenterà ora di dare una lettura più approfondita dei dati riportati nella sezione precedente, insieme ad alcune riflessioni finali.

L'esperimento qui illustrato ha dimostrato che la scala elaborata da Kuiken e Vedder (2017) presa in esame può essere considerata uno strumento più che affidabile per la misurazione dell'adeguatezza funzionale. Lo strumento sembra essere ben tarato rispetto agli obiettivi prefissati: misurare oggettivamente le dimensioni pragmatiche del linguaggio e la felice riuscita del compito.

L'analisi statistica ha infatti dimostrato che l'*interrater reliability*, impiegata come misura dell'affidabilità della scala, raggiunge livelli buoni o molto buoni per le singole dimensioni del costrutto; ciò prova che i descrittori sono formulati piuttosto chiaramente.

Allo stesso tempo la valutazione dell'*interrater agreement* ha dato risultati più che soddisfacenti, dimostrando che anche in termini assoluti vi è un sostanziale accordo tra i valutatori; tale elemento testimonia, ancora, la precisione della formulazione dei singoli livelli di ogni scala. La misurazione, infatti, registra che i giudizi assegnati dai 4 valutatori non si discostano sensibilmente e in molti casi, anzi, possono coincidere.

Per quanto concerne le correlazioni tra le quattro aree dell'adeguatezza funzionale, va sicuramente segnalata l'elevata relazione tra contenuto e requisiti del task in tutti i tipi di testi, con valori veramente buoni; ciò evidentemente si può spiegare con il fatto che entrambe le dimensioni siano state interpretate in modo spiccatamente quantitativo dai valutatori. La dimensione del contenuto è valutata (così come formulata dai descrittori) in termini accentuatamente quantitativi (si fa esplicito riferimento alla misura del numero delle unità informative nei testi); le consegne dei singoli task, invece, prevedevano una serie di domande e obiettivi da soddisfare per costruire il proprio testo. Le interviste retrospettive hanno indicato che i quattro *rater* hanno spesso fatto riferimento a tali istruzioni come obiettivi per considerare riuscito il compito, che prevedeva, tra l'altro, anche una lunghezza minima di parole. L'interazione di questi fattori può aver prodotto, quindi, una stretta correlazione tra le due dimensioni.

Così come in Vedder (2016), in cui il valore era di ,938, si registra una forte correlazione anche tra coerenza/coesione e comprensibilità del testo, a prescindere dal task; il che forse può non stupire: un testo coerente e coeso è anche un testo ampiamente comprensibile, in cui i legami semantici e sintattico-testuali sono esplicitati con chiarezza.

I dati rilevati nei primi esperimenti proposti dagli ideatori delle scale (Vedder, 2016, Kuiken, Vedder, 2017; 2018) mostravano una correlazione meno marcata (,544) rispetto alle altre tra la dimensione dei requisiti del task e della comprensibilità e della coerenza/coesione. Nel presente studio qualcosa di simile può essere osservato invece solo per quanto riguarda il task narrativo tra i requisiti del task e la comprensibilità; si tratta, ad ogni modo, di correlazioni decisamente sufficienti.

In sintesi, questi dati testimoniano la bontà della teorizzazione del costrutto come multidimensionale e nello specifico nella sua composizione in quattro aree da valutare separatamente, data anche la netta differenza teorico-concettuale tra di esse. La correlazione tra le dimensioni metterebbe, insomma, in evidenza come i costrutti sottostanti interagiscano positivamente tra loro e possano essere considerati validi per la definizione complessiva dell'adeguatezza funzionale.

I dati, in definitiva, sembrano però puntare anche verso qualche difficoltà nell'applicazione delle scale, soprattutto per la dimensione di coerenza/coesione, così come emerso dalle fasi di addestramento all'utilizzo dello strumento e nelle interviste retrospettive. I valutatori, infatti, hanno sottolineato a più riprese la complessità del trovare un equilibrio nella giusta valutazione tra coerenza e coesione: se per la prima parte del costrutto un livello accettabile appare raggiungibile in molti casi (non sono molti i salti logici o argomentativi), la seconda risulta assai più complessa, soprattutto per l'assenza molto spesso di elementi coesivi testuali. D'altro canto, la formulazione dei descrittori in questa dimensione palesa l'originaria destinazione della scala, pensata e applicata a task di tipo argomentativo, nei quali è lecito attendere una maggiore ricchezza di forme coesive.

L'ultimo aspetto valutato è quello della correlazione tra la competenza linguistica degli studenti e l'adeguatezza funzionale e le sue dimensioni. Vi è una correlazione sufficiente tra i risultati del C-test e la valutazione globale del costrutto per tutti e tre i testi, ma non tale da rendere la competenza linguistica generale una prerogativa necessaria per il soddisfacimento, seppur minimo, dei compiti; ciò sembrerebbe suggerire che il costrutto può essere valutato a prescindere da indici puramente linguistici (come quelli CAF).

Anche se, come accennato in precedenza, è necessaria una certa cautela nell'interpretazione del dato, certamente da non sovrastimare.

Relativamente alle singole dimensioni, i risultati variano leggermente in dipendenza dal compito: il task argomentativo è l'unico a mostrare correlazioni significative in tutte e quattro le dimensioni. Nel task regolativo e in quello narrativo la relazione tra il contenuto e la competenza linguistica invece non appare sufficiente; così come il valore dei requisiti del task nel testo narrativo.

Dai commenti forniti dai valutatori, sia durante il *training* sia nelle interviste retrospettive, possiamo ricavare alcune motivazioni. Per quanto concerne il contenuto, l'impostazione spiccatamente quantitativa data al costrutto e il fatto che nei task fossero presenti delle linee guida precise che aiutassero a costruire il proprio testo ha fatto sì che gli informanti riuscissero a produrre un numero di idee per lo meno sufficiente a rispondere alle richieste nonostante un livello linguistico non elevato. La tipologia testuale è un altro elemento addotto dai valutatori: il testo narrativo è stato il più problematico da valutare. Secondo i *rater*, infatti, la sensazione è che spesso, a prescindere dall'effettiva comprensibilità o correttezza linguistica dei testi, gli informanti non riuscissero ad aderire al contesto situazionale (il concorso di scrittura) e allo specifico atto linguistico della narrazione.

La scala si è dimostrata uno strumento affidabile; il suo utilizzo è stato abbondantemente coerente tra i quattro valutatori, che seppur non mostrando un accordo assoluto, hanno comunque compreso i descrittori e li hanno applicati in modo simile.

Le correlazioni significative, tra il sufficiente e il molto buono, registrate tra le varie scale dimostrano che le quattro dimensioni individuate come parte del costrutto globale interagiscono positivamente tra di loro e insieme contribuiscono alla valutazione finale dell'adeguatezza funzionale. La scelta di adoperare quattro scale differenti è comunque indispensabile sia considerando l'ampia differenza teorico-concettuale tra i singoli costrutti, sia per misurare in maniera oggettiva e dettagliata tali aspetti.

Il presente studio dimostra ancora che la scala può essere utilizzata anche da valutatori non esperti; la sostanziale coerenza tra le valutazioni sicuramente fra propendere per un giudizio positivo dello strumento.

Tuttavia, i commenti dei valutatori hanno messo in luce alcune possibili criticità nella formulazione di alcuni descrittori nella combinazione tra le due dimensioni di coerenza e coesione. I quattro *rater* hanno sottolineato come sia a volte difficile giungere a un singolo punteggio che tenga bilanciati i due aspetti testuali: se da un lato gli informanti potevano raggiungere punteggi tutto sommato buoni a livello di coerenza (con assenza di salti logici evidenti), dall'altro lato, assai più difficoltosa si rivelava la valutazione degli elementi coesivi, data la generale assenza di un ampio numero di connettivi di varia natura in gran parte dei testi. Tale aspetto è inoltre sensibile allo specifico compito comunicativo. Nel caso, ad esempio, del task regolativo (stesura di un biglietto con delle istruzioni per gli ospiti) appare piuttosto improbabile la comparsa nel testo di forme coesive complesse. A tal proposito, è bene ricordare che la scala è stata sviluppata in principio pensando a testi argomentativi; ciò ha ovvie ricadute sulla formulazione dei descrittori: si pensi, ad esempio, al focus sulla quantità e la coerenza delle idee della dimensione del contenuto o sull'insistenza sugli aspetti coesivi per la scala della coerenza/coesione. Sono questi aspetti assai sensibili al variare della tipologia testuale; la scala presenta senza dubbio alcuni limiti nell'applicazione in tal senso. I descrittori potrebbero in tal senso essere ancora affinati.

I dati raccolti e analizzati costituiscono un campione interessante, anche se ancora da espandere, pensando, ad esempio, a raccogliere anche un *corpus* parallelo di produzioni orali sugli stessi informanti. Lo strumento valutativo proposto e analizzato in questa sede

si è dimostrato utile; l'obiettivo di testare la scala su tre diverse tipologie di task oltre a quella argomentativa ha dato buoni risultati, ma allo stesso tempo indica alcuni passi per un progressivo affinamento dei descrittori che ne amplino l'applicabilità.

RIFERIMENTI BIBLIOGRAFICI

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

Bridgeman B., Powers D., Stone E., Mollan P. (2012), "TOEFL iBT speaking test scores as indicators of oral communicative language proficiency", in *Language Testing*, 29, 1, pp. 91-108.

Council of Europe (2001), *Common European Framework for Languages: Learning, Teaching, Assessment*, Strasbourg. Trad. It. (2002), *Quadro comune europeo di riferimento per le lingue*: *apprendimento*, *insegnamento*, *valutazione*, La Nuova Italia, Oxford, Firenze.

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

Del Bono F. (2019), "Aspetti pragmatici nella valutazione di testi scritti: uno studio sull'adeguatezza funzionale in italiano L2", in Nuzzo E., Vedder I. (a cura di), *Lingua in contesto*: *la prospettiva pragmatica*, Studi AItLA, Officinaventuno, Milano, pp. 231-244: http://www.aitla.it/images/pdf/StudiAItLA9/014_DelBono.pdf.

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, pp. 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, pp. 121-142.

Faone S., Pagliara F. (in stampa), "How to assess L2 information-gap tasks through Functional Adequacy rating scales", 7th TBLT Conference Tasks in Context, University of Barcelona, Barcellona, April 19-21.

Fragai E. (2001), "La programmazione didattica: Il Glotto-Kit come strumento per valutare i livelli in entrata", in Barni M., Villarini A. (a cura di), *La questione della lingua per gli immigrati stranieri. Insegnare*, *valutare e certificare l'Italiano L2*, FrancoAngeli, Milano, pp.191-208.

Grice H. P. (1975), "Logic and conversation", in Cole P., Morgan J. L. (eds.), *Speech acts*, Academic Press, New York, pp. 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, pp. 805-817.

Knoch U. (2009), "Diagnostic assessment of writing: A comparison of two rating scales", in *Language Testing*, 26, 2, pp. 275-304.

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

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

Kuiken F., Vedder I. (2018), "Assessing functional adequacy of L2 performance in a task-based approach", in Taguchi N., YouJin K. (eds.), *Task-Based Approaches to Teaching and Assessing Pragmatics*, John Benjamins, Amsterdam, pp. 266-285.

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*: *Instersections between SLA and language testing research*, EuroSLA Monograph Series 1, Amsterdam, pp. 81-100.

McNamara T., Roever C. (2007), *Testing*: *The social dimension*, Blackwell Publishing, Malden, MA/Oxford UK.

Nuzzo E., Santoro E. (2017), "Apprendimento, insegnamento e uso di competenze pragmatiche in italiano L2/LS: la ricerca a partire dagli anni Duemila", in *EuroAmerican Journal of Applied Linguistics and Languages*, 4, 2, pp. 1-27.

Pagliara F. (in stampa), "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, pp. 590-601.

Révész A., Ekiert M., Torgersen E. N. (2016), "The Effects of Complexity, Accuracy, and Fluency on Communicative Adequacy in Oral Task Performance", in *Applied Linguistics*, 37, 6, pp. 828-848.

Sato T. (2012), "The contribution of test-takers' speech content to scores on an English oral proficiency test", in *Language Testing*, 29, 2, pp. 223-241.

Upshur J. A., Turner C. E. (1995), "Constructing rating scales for second language tests", in *ELT Journal*, 49, 1, pp. 3-12.

APPENDICE

Contenuto: il numero delle unità informative espresse nel testo è adeguato e rilevante?

1. *Per niente*: il numero di idee è insufficiente, non è affatto adeguato, e non sono coerenti le une con le altre.
2. *Appena*: il numero di idee non è sufficientemente adeguato, le idee sono poco consistenti.
3. *Parzialmente*: il numero di idee è abbastanza adeguato anche se non sono molto consistenti.
4. *Sufficientemente*: il numero di idee è adeguato e sono sufficientemente consistenti.
5. *Largamente*: Il numero di idee è molto adeguato e appaiono molto consistenti e coerenti le une con le altre.
6. *Assolutamente*: il numero di idee è assolutamente adeguato e appaiono molto consistenti e coerenti le une con le altre.

Requisiti del task: Sono stati soddisfatti i requisiti e i criteri specifici del task (ad es. genere, atto linguistico, registro)?

1. *Per niente*: nessuno dei requisiti e dei criteri specifici del task è stato soddisfatto.
2. *Appena*: pochi requisiti e criteri specifici del task (meno della metà) sono stati soddisfatti.
3. *Parzialmente*: circa la metà dei requisiti e dei criteri specifici del task sono stati soddisfatti.
4. *Sufficientemente*: più della metà dei requisiti e delle condizioni del task sono stati soddisfatti.
5. *Largamente*: quasi tutti i requisiti e i criteri specifici del task sono stati soddisfatti.
6. *Assolutamente*: tutti i requisiti e i criteri specifici del task sono stati soddisfatti.

Comprensibilità: quanto sforzo è richiesto per capire lo scopo del testo e le idee?

1. *Per niente*: il testo non è affatto comprensibile. Le idee e lo scopo sono espressi in modo oscuro e il lettore, anche sforzandosi, non riesce a capire.
2. *Appena*: il testo non si comprende facilmente, i suoi scopi non sono chiari e il lettore deve sforzarsi molto per capire le idee dell'autore. Il lettore deve cercare di indovinare la maggior parte delle idee e degli scopi del testo.
3. *Parzialmente*: il testo è parzialmente comprensibile, ma alcune frasi non si capiscono bene a una prima lettura. Un'ulteriore rilettura è utile per chiarire gli scopi del testo e le idee espresse, ma rimangono alcuni dubbi.
4. *Sufficientemente*: il testo è sufficientemente comprensibile, solo certe frasi sono poco chiare ma si possono capire senza grandi sforzi con una rilettura.
5. *Largamente comprensibile*: il testo è facile da comprendere e si legge agevolmente, non ci sono problemi di comprensibilità.
6. *Assolutamente*: il testo è molto facile da comprendere e si legge molto agevolmente, le idee e gli scopi sono espressi con chiarezza.

Coerenza e coesione: il testo è coerente e coeso (ad es. ci sono connettivi e strategie di coesione)?

1. *Per niente*: ci sono frequenti salti logici e argomenti non collegati. L'autore non usa riferimenti anaforici (pronomi, frasi con soggetto sottinteso chiaramente interpretabile). Il testo non è affatto coeso. I connettivi sono praticamente assenti e le idee non sono collegate tra loro.
2. *Appena*: l'autore spesso non collega gli argomenti tra loro; se c'è coerenza, questa è espressa per lo più da ripetizioni. Pochi riferimenti anaforici. Alcuni salti logici. Il testo è poco coeso. Vengono usati pochi connettivi, che non collegano bene le idee.
3. *Parzialmente*: ci sono frequenti salti di argomento e/o ripetizioni. Più di due frasi di seguito esprimono esplicitamente lo stesso soggetto, anche quando questo sarebbe chiaro. Vengono usati alcuni riferimenti anaforici. Possono esservi alcune interruzioni della coerenza. Il testo è abbastanza coeso. Vengono usati alcuni connettivi, ma sono per lo più delle congiunzioni.
4. *Sufficientemente*: i salti di argomento sono abbastanza rari, ma l'autore a volte riesce a essere coerente solo mediante ripetizioni non necessarie. Si trovano sufficienti riferimenti anaforici. Possono esservi interruzioni della coerenza. L'autore fa un buon uso dei connettivi, che a volte vanno oltre le semplici congiunzioni.
5. *Largamente*: quando si introduce un nuovo argomento, di solito ciò avviene mediante l'uso di connettivi o espressioni di collegamento esplicite. Le ripetizioni sono molto rare. Si trovano numerosi riferimenti anaforici. Nessuna interruzione della coerenza. Il testo è molto coeso e le idee sono ben collegate tra loro mediante connettivi nominali o avverbiali.
6. *Assolutamente:* l'autore produce un'ottima coerenza integrando le nuove idee nel testo con connettivi o espressioni di collegamento esplicite. I connettivi anaforici sono usati regolarmente. Qualche raro caso di argomenti non collegati e nessuna interruzione della coerenza. La struttura del testo è estremamente coesa, grazie a un abile uso dei connettivi (in particolare nominali, avverbiali e formule di collegamento), usati spesso per descrivere le relazioni tra le idee.