DOI: 10.17469/O2108SLI000014

SIMONE MATTIOLA, EMANUELE MIOLA

Nuove risposte a vecchie domande: considerazioni sul *continuum* italo-romanzo alla luce dell'ASJP

Il presente contributo mira a presentare l'ASJP, un nuovo metodo lessico-fonetico che permette di calcolare la distanza strutturale tra oggetti linguistici differenti e pertanto, tramite l'identificazione di una soglia di omogeneità, distinguere quando essi sono da considerarsi come lingue differenti e quando invece come varietà di una stessa lingua. Questo metodo verrà poi applicato e testato sul cosiddetto *continuum* linguistico gallo-italico del Nord Italia andando a calcolare i valori di omogeneità tra gli oggetti selezionati. In questo modo potremo osservare e discutere il loro statuto linguistico all'interno del panorama linguistico italiano settentrionale e provare quindi a contare il numero di lingue che in esso possiamo identificare. Infine, concluderemo discutendo i punti di forza e i punti deboli dell'ASJP proponendo anche qualche possibile spunto per il futuro.

Parole chiave: lingue, varietà di lingua, distanza strutturale, ASJP.

1. *Introduzione*

Questo contributo deve necessariamente iniziare con un'*explicatio terminorum*. Mentre *oggetto linguistico* verrà qui usato con il significato di qualsiasi realizzazione della capacità di parola che può presentarsi a chi fa ricerca linguistica (ma anche al semplice ascoltatore) come un codice scritto o parlato o segnato atto a comunicare, impiegheremo *lingua* (o *sistema linguistico*, usato come sinonimo *tout court* di *lingua*) nel significato di oggetto linguistico adoperato, dai suoi parlanti, in modi differenti (in diafasia, diamesia, diastratia e diatopia), ma comunque tra loro altamente comprensibili. Questi modi, a loro

volta, saranno indicati con il termine *varietà* (o *varietà della lingua x*, poniamo *varietà dell'inglese*)[1].

I concetti in questione vengono poi osservati qui in una prospettiva che chiameremmo linguistico-teorica. Non vengono osservati, cioè, da un punto di vista strettamente repertoristico e variazionistico, come può essere quello di chi si occupa, specialmente muovendosi nell'alveo della tradizione europea, di dialettologia e di sociolinguistica. Queste ultime due branche, infatti, usano le etichette di *lingua* e *varietà* applicandovi talvolta definizioni in tutto o in parte diverse da quelle che abbiamo dato noi, perché sono definizioni specifiche di quegli ambiti di indagine che sono votati a studiare la diversità verticale, ovvero interna alle lingue (Grandi 2020). Per quegli ambiti di indagine, tali definizioni possiedono una chiara connotazione scientifica e capacità euristica. Ciononostante, non di rado l'impiego delle due etichette persino in queste branche può essere poco coerente anche nelle pubblicazioni scientifiche: Berruto (2009: 342-343) notava che da un punto di vista squisitamente sociolinguistico

> non ha fondamento considerare il friulano e il sardo effettive lingue minoritarie, o, se li si considera tali, bisognerebbe trattare come lingue minoritarie anche i dialetti piemontese, lombardo, veneto, napoletano, siciliano e così via,

riducendo dunque, pure nei lavori di taglio eminentemente sociolinguistico, a due i tre *entia* lingua, lingua minoritaria e dialetto, che sono invece di larghissimo uso nei nostri lavori. Per questi motivi, ma non unicamente per questi, non è forse peregrino invocare la necessità di una terminologia univoca almeno nelle diverse branche della stessa disciplina, foss'anche di una disciplina come la linguistica, ovvero di una scienza non-dura.

La questione di che cosa sia una lingua e di che cosa sia una varietà di lingua o, per meglio dire, di quando due oggetti linguistici

[1] Il concetto di varietà di lingua così definito rassomiglia molto da vicino a quello dell'inglese *dialect*, almeno nei suoi impieghi maggioritari nei lavori che si occupano di linguistica teorica. *Dialect*, o la sua traduzione automatica nonché falso amico *dialetto*, non saranno adoperati al fine di evitare prevedibili incomprensioni. Anche *varietà*, in linguistica, è comunque ben lontano da quella monosemia referenziale che dovrebbe essere propria dei termini scientifici, dato che può indicare sia oggetti linguistici come l'inglese sia oggetti linguistici come l'inglese di Londra o di New York, benché sia evidente che le ultime due "varietà" sono ricomprese all'interno del primo.

siano due lingue diverse o due varietà della stessa lingua, non è invece ininfluente per quelle branche della linguistica che debbono badare con maggiore attenzione alla diversità orizzontale, cioè alle differenze che corrono tra lingue diverse. Chi fa tipologia, ad esempio, ambendo a catturare la *diversitas linguarum* deve avere ben chiaro se l'oggetto linguistico che sta osservando è una lingua a sé oppure una varietà di un altro degli oggetti linguistici parlati in quella stessa comunità linguistica o nelle comunità circonvicine. Si pensi alla condizione che si incontra in Lituania: il lituano ha SVO come ordine non marcato dei costituenti nella frase dichiarativa, ma "[n]ella lingua popolare e nel folklore l'ordine (S)OV è più frequente" (Schmalstieg 1993: 503; v. anche Mathiassen 1996: 240-241; e cfr. Ambrazas 1997: 695). Di fronte a una differenza strutturale tanto marcata è ragionevole che chi fa tipologia si domandi se si trovi di fronte a una lingua sola, di cui osserva due varietà diverse, oppure a due lingue differenti. Rispondere non è per nulla scontato, dal momento che è risaputo che le varietà di una stessa lingua possono avere profili tipologici anche molto diversi (cfr. Grandi 2018 per l'italiano neo-standard in relazione allo standard), e la risposta, quale che sia, ha risvolti che possono essere anche importanti sulla ricerca. Sembra quindi utile l'elaborazione di un criterio scientificamente rigoroso e formalizzabile. Ma la nostra riflessione parrà non inutile anche a chi si occupa di pianificazione linguistica, per diversi motivi, ma primariamente perché, dato che la pianificazione ha lo scopo di preservare la diversità orizzontale, deve concentrare il suo impegno nel tutelare la gamma di varietà che costituiscono una lingua minoritaria o minorizzata, e non nel tutelare come lingua minoritaria una varietà di una lingua ufficiale. Anche in questo caso la necessità di comprendere bene e in modo chiaro che cosa sia lingua e che cosa sia varietà pare ineludibile.

Ma soprattutto, a nostro modo di vedere, la questione non è ininfluente dal punto di vista scientifico-speculativo: se una domanda di ricerca permane o non ha trovato ancora risposta unanime da parte della comunità scientifica, è lecito a studiose e studiosi porla nuovamente, specie – com'è nel nostro caso (v. §2) – in presenza di nuovi metodi e di nuove tecnologie disponibili per l'indagine.

Del resto, del problema di un'accurata definizione dei termini tecnici della nostra disciplina si sta occupando da tempo Martin Haspelmath (2013, 2021, 2023, tra i molti altri) né sono mancate anche in anni re-

centi proposte intorno al problema specifico che ci proponiamo di trattare, come quella di Cysouw & Good (2013). Anche questi due autori prendevano le mosse dalla considerazione che "[t]he fact that people often use the same names with rather different intentions is a fact of life in informal discourse" mentre "one of the distinguishing features of scholarly communication is attending to this problem when it may cause crucial misunderstandings"[2], ma si rivelavano scettici sulla possibilità di una "definition of the term 'language' that would satisfy all users and at the same time be scientifically rigorous" (Cysouw & Good 2013: 331)[3].

Di altro avviso sembra essere Søren Wichmann. Questi ritiene che la questione di come distinguere lingue diverse da varietà diverse di una stessa lingua sia "useful for many different purposes" in linguistica,

> [m]ore importantly perhaps, if such a distinction is a feature of the way that language varieties are distributed rather than just a distinction we impose in some arbitrary way, then this would be important for the understanding of the sociology of language at large. (Wichmann 2019: 1)

Wichmann ritiene anche di poter distinguere lingue diverse da varietà di una stessa lingua misurando la distanza linguistica tra due diversi oggetti linguistici tramite un nuovo metodo computazionale su cui si basa il progetto ASJP (*Automated Similarity Judgment Program*), pubblicato online sotto forma di database (https://asjp.clld.org/, Wichmann *et al.* 2022).

L'obiettivo di questo nostro contributo consiste nel presentare e nell'applicare questo metodo al *continuum* linguistico dell'Italia settentrionale per identificare i possibili confini linguistici, ma anche per

---

[2] Non è nemmeno il caso di additare i fraintendimenti a cui un impiego non coerente delle etichette *lingua* e *varietà* (e/o *dialetto*) può dare luogo sia nella ricerca scientifica, specie internazionale, sia nell'istruzione, a tutti i gradi (cfr. Cravens 2022; Miola 2020).

[3] La proposta terminologica avanzata da Cysouw & Good (2013) non sembra soddisfacente, nemmeno per gli scopi sommariamente esposti poc'anzi nel corpo del testo. Se, infatti, *doculect* (o *documented lect*) vale "a linguistic variety as it is documented in a given resource" (Cysouw & Good 2013: 342) e può sovrapporsi abbastanza bene alla nostra definizione di oggetto linguistico, *languoid*, ovvero "any (possibly hierarchical) grouping of doculects", è un termine che ha un'estensione troppo ampia, dato che può applicarsi "from a set of idiolects to a high-level language family" (Cysouw & Good 2013: 347): in pratica l'italiano può essere considerato un singolo *languoid*, ma anche le lingue romanze possono esserlo.

verificare la solidità metodologica dell'ASJP testandolo su un territorio linguisticamente complesso, come quello italiano, fatalmente sempre più trascurato – forse perché non conosciuto a fondo – dagli studi internazionali.

Più nello specifico, quindi, si calcoleranno con il metodo lessico-statistico dell'ASJP le distanze tra diversi oggetti linguistici parlati nell'Italia settentrionale (§3) con il fine di valutare i risultati ottenuti mettendo in luce eventuali problemi e possibili miglioramenti dell'ASJP (§4). Prima, però, occorrerà inquadrare all'interno degli studi pregressi l'ASJP e spiegarne almeno sommariamente il metodo (§2).

2. *Il metodo ASJP*

L'ASJP non è certamente il primo metodo proposto per calcolare le distanze linguistiche, ma si inserisce in una tradizione lunga e ben solida, almeno per quanto riguarda la dialettologia italiana e romanza.

Tra i tanti studi condotti con l'intento di trovare una classificazione delle lingue romanze sulla base delle loro distanze strutturali ci piace qui ricordare almeno Pei (1949), Grimes & Agard (1959), Muljačić (1967), Iliescu (1969), Devoto (1970), Pellegrini (1970) e Francescato (1980). Questi meritori lavori, fondamentali per dissodare il campo di ricerca, soffrono però di un certo numero di problemi metodologici. Innanzitutto, la selezione dei tratti rilevanti viene operata a monte da chi si appresta a studiare la situazione, procedimento che, non essendo esente da circolarità di ragionamento, rischia di condizionare fortemente i risultati, per esempio quando si tratta di individuare raggruppamenti di lingue (v. Francescato 1972). Scegliendo un determinato mazzetto di tratti, per fare un esempio, sarebbe possibile mostrare che tra italiano e francese non vi sono differenze strutturali. Inoltre, "viene evidentemente trascurato con questo metodo [...] il fatto che non tutti i tratti [...] hanno [...] la stessa rilevanza" dal punto di vista linguistico (Grassi *et al.* 2001: 79). Un tratto che può sembrare rilevante dal punto di vista della fonetica storica, come per esempio quello dell'esito scempio o non scempio della geminata del latino *stēlla* (tratto 26 di Muljačić 1967 ripreso da Pellegrini 1970), intuitivamente distanzia due oggetti linguistici decisamente meno di quanto possa farlo un tratto

morfosintattico, come la formazione del plurale dei nomi (tratto 13) o del futuro (tratto 20). Infine, resta il fatto che i confini tra lingue diverse vanno tracciati "in corrispondenza di quei punti in cui la differenza tra grammatiche aumenta in modo significativo" (Sobrero 1993: 9), ma non è generalmente mai chiarita quale sia la differenza-soglia che permette di parlare di (grammatiche di) lingue diverse. Ciò non toglie che le indagini sistematiche tramite il metodo lessico-statistico restano quelle invocate da più parti come probabilmente capaci di portare, a questo riguardo, risultati utili (v. Berruto 1993: 3n).

Il metodo dell'ASJP offre una risposta a tutti i vari ordini di problemi elencati poc'anzi e lavora, con approccio computazionale, proprio attraverso la lessico-statistica[4]. Wichmann (2019) ha infatti proposto un metodo computazionale che va a misurare la distanza linguistica tra oggetti linguistici e ha anche individuato la soglia di omogeneità (*sameness* nelle sue parole) oltre la quale due oggetti linguistici sarebbero da considerarsi lingue differenti e al di sotto della quale essi sarebbero da considerarsi due varietà di una stessa lingua. Questo approccio è alla base del già citato progetto ASJP (Wichmann *et al.* 2022). Questo database contiene liste di 40 parole (corrispondenti a un sottoinsieme delle 100 parole che compongono la lista di Swadesh, v. Swadesh 1950) del vocabolario di base per 7.655 doculetti (*doculects)*, i quali rappresentano circa i due terzi del totale degli oggetti linguistici esistenti sul nostro pianeta. Le liste sono trascritte secondo l'ASJPcode (Brown *et al.* 2013: 7-9), una sorta di grafia standardizzata fatta di convenzioni che, partendo dalla trascrizione IPA, sono volte a ridurre le differenze prevedibili foneticamente e fonologicamente.

Per il metodo di calcolo di distanza strutturale alla base dell'ASJP sono stati utilizzati solo i doculetti per cui era a disposizione almeno il 70% delle parole presenti nella lista di 40 (28/40), e questo perché, secondo il test dell'alpha di Cronbach (Jäger 2015), si tratta della soglia sopra la quale i calcoli risultano essere affidabili per scopi filogenetici. Pertanto, il conto finale è di 5.800 doculetti.

[4] Anche la dialettometria, come si vedrà con metodo non completamente dissimile da quello dell'ASJP, si è occupata dei problemi esposti in questo paragrafo e della loro risoluzione (v. Goebl 1987, 2008, *inter alia*; un'applicazione della dialettometria alle varietà gallo-italiche è stata presentata da Tamburelli & Brasca 2018).

La distanza tra due oggetti linguistici viene misurata tramite una versione creata *ad hoc* della distanza di Levenshtein (d'ora in avanti LD), che calcola il numero di sostituzioni, inserzioni o cancellazioni necessarie per trasformare una parola in un'altra. Quindi: date due liste di parole di due oggetti linguistici, si può misurare la LD per ciascuna coppia di lessemi corrispondenti nella lista di 40 parole e dividerla per la lunghezza della parola più lunga della coppia andando a identificare quello che viene chiamata distanza di Levenshtein normalizzata (d'ora in poi LDN). In questo modo, si avrà un valore numerico compreso tra 0 e 1, i poli del *continuum* di omogeneità, che rappresentano rispettivamente la situazione in cui due oggetti linguistici sono due varietà della stessa lingua (LDN più vicino a 0) e la situazione in cui due oggetti linguistici sono due lingue totalmente differenti (LDN = 1), cioè senza alcun tipo di relazione lessico-fonetica tra tutte le coppie di parole. Sul *continuum* tra 0 e 1, che ci dà solamente la distanza tra due oggetti linguistici senza dirci nulla sul loro reale status, bisogna, però, identificare una soglia di confine (detta soglia di *cut-off*) oltre la quale due oggetti linguistici siano da considerarsi due lingue diverse e sotto la quale, invece, essi siano da considerarsi come due varietà della medesima lingua. Ma come identificare il valore soglia in maniera oggettiva, cioè senza adottare un valore di *cut-off* che sia arbitrariamente deciso dal ricercatore? Per rispondere a questa domanda, Wichmann ha calcolato il valore di LDN tra tutti gli oggetti linguistici noti all'interno di 15 *genera* linguistici differenti e accuratamente selezionati tra quelli per i quali il noto database Ethnologue (Simons & Fennig 2017) mostrasse almeno il 10% di oggetti linguistici con il medesimo codice ISO 639-3 (ovvero oggetti considerati dal punto di vista qualitativo come varietà della medesima lingua). Wichmann (2019) ha quindi analizzato le distribuzioni dei vari valori di LDN tra oggetti linguistici con medesimo codice ISO 639-3 e tra oggetti linguistici con codici ISO 639-3 differenti tramite una serie di tecniche e test di natura computazionale e statistica, che per questioni di spazio non possiamo approfondire (rimandiamo direttamente a Wichmann 2019: 826-828 per la descrizione e l'analisi di queste tecniche). La soglia di *cut-off* finale risultante dall'analisi delle varie distribuzioni di LDN dei 15 *genera* è di 0,51 (approssimazione di

0,5138 ± 0,0707)[5]. Pertanto: due oggetti linguistici che presentano un valore di LDN che si posiziona al di sotto della soglia di *cut-off* 0,51 saranno molto probabilmente da considerarsi come due varietà della medesima lingua, mentre, al contrario, due oggetti linguistici che presentano un valore di LDN che si posiziona al di sopra della soglia di *cut-off* 0,51 saranno molto probabilmente da considerarsi come due lingue separate.

Wichmann (2019) ha poi provato ad applicare questo metodo e la relativa soglia di *cut-off* identificata in una serie di situazioni ritenute piuttosto complesse. I valori di LDN tra gli oggetti linguistici calcolati da Wichmann (2019: 829) sono riportati in Tabella 1.

Osservando questa tabella possiamo fare alcune rilevanti considerazioni. Innanzitutto, è da notare come, in realtà, la distanza tra gli oggetti linguistici sia spesso rappresentata da un valore di LDN che difficilmente si attesta vicino allo 0,51 (anche nei casi pensati come di "confine"). Infatti, ad esclusione delle coppie danese/svedese e groenlandese orientale/groenlandese occidentale, che si posizionano molto vicine alla soglia di *cut-off*, le restanti coppie di oggetti linguistici hanno un valore LDN abbastanza distante da 0,51. Ovviamente, non mancano alcune situazioni "ambigue", ma sono meno di quelle che ci si poteva aspettare. Pertanto, secondo questi dati, bosniaco e croato sarebbero due varietà della stessa lingua, così come lo hindi e l'urdu, ma anche il tedesco standard e il tedesco di Berna e il russo e il bielorusso (sebbene con una differenza più marcata). Al contrario, l'arabo parlato in Egitto e quello marocchino sono da considerarsi come due lingue diverse, così come la varietà cinese di Dongshan e quella di Fuzhou e come il catalano e lo spagnolo. Solamente le coppie citate all'inizio di questo capoverso (danese/svedese e groenlandese orientale/groenlandese occidentale) rappresentano una situazione di reale ambiguità. Il metodo dell'ASJP sembra quindi aiutare, e di molto, il difficile compito di distinguere rigorosamente e scientificamente due oggetti linguisti-

[5] È importante sottolineare come Wichmann (2019: 828) ritenga che la soglia di *cut-off* sia probabilmente da posizionarsi leggermente sopra quella proposta (cioè 0,5686 ± 0,0072), sulla base di ulteriori calcoli preliminari da lui stesso operati, ma per ragioni di solidità metodologica e teorica ha ritenuto più opportuno adottare la soglia più conservativa in quanto più affidabile, almeno al presente livello di analisi.

ci come varietà di una stessa lingua o due lingue separate anche in quelle situazioni che la letteratura descrive come più ambigue.

Tabella 1 - *Valori LDN per coppie di oggetti linguistici dibattuti in letteratura sulla questione lingua/varietà di lingua (Wichmann 2019: 829)*

Oggetto linguistico A	**Oggetto linguistico B**	**LDN**
indonesiano	malese	0,1199
bosniaco	croato	0,1324
quechua di Chachapoyas	quechua di Huaylas Ancash	0,3055
hindi	urdu	0,4281
nahuatl classico	pipil	0,4336
tedesco standard	tedesco di Berna (svizzero)	0,4638
russo	bielorusso	0,4647
danese	svedese	0,4921
groenlandese orientale	groenlandese occidentale	0,5036
navajo	apache jicarilla	0,5708
arabo del Cairo (egiziano)	arabo marocchino	0,5814
cinese di Dongshan	cinese di Fuzhou	0,6013
catalano	spagnolo	0,6589
giapponese	miyako (ryukyuano)	0,6680

3. *Applicazione dell'ASJP alla situazione linguistica del Nord Italia*

Come già detto nell'introduzione, il nostro obiettivo principale è quello di applicare il metodo del progetto ASJP al *continuum* linguistico dell'Italia settentrionale in modo tale da identificare i possibili confini linguistici e testare la solidità metodologica dell'ASJP su un territorio linguisticamente complesso e poco conosciuto, almeno ne-

gli studi internazionali. Adotteremo come soglia di *cut-off* 0,51 così come proposto da Wichmann (2019)[6].

Abbiamo raccolto da colleghi parlanti fluenti e/o esperti le liste di parole, trascritte secondo l'IPA, per una serie di oggetti linguistici distribuiti nell'area settentrionale d'Italia in modo tale da avere una buona copertura geografica e linguistica. Gli oggetti linguistici per i quali abbiamo raccolto le liste di parole sono: il piemontese torinese (codice ISO 639-3 PMS), il lombardo bergamasco (LMO), il lombardo pavese (LMO), il veneto veneziano (VEC), l'emiliano bolognese (EGL) e il romagnolo forlivese (RGN). Come già parzialmente notato, la scelta di questi oggetti non è casuale e la si può ricondurre a due criteri di base: (a) cercare di rappresentare il maggior numero di zone geografiche del Nord Italia (per quanto possibile), (b) selezionare oggetti per i quali potevamo avere dati di prima mano e soprattutto affidabili (questo perché, come si vedrà nel §4, alcune liste dell'ASJP non lo sono). A questi abbiamo aggiunto altri oggetti "di controllo", più e meno vicini geo-linguisticamente e tipologicamente: italiano standard (ITA), portorecanatese (ITA), siciliano palermitano (SCN), francese (FRA), catalano (CAT), giapponese (JPN). Abbiamo quindi calcolato tramite il software ASJP i valori di LDN per ciascuno di questi oggetti in relazione agli altri della lista. I risultati sono riportati in Tabella 2[7].

[6] Per comodità argomentativa, non presenteremo i vari valori di LDN (compreso il valore *cut-off*) con i relativi range di variazione statistica dovuti al margine di errore.

[7] È importante notare che spesso abbiamo volutamente selezionato varietà cittadine degli oggetti linguistici analizzati e che, pertanto, è prevedibile una sensibile riduzione del valore di LDN tra questi e l'italiano (a causa dei fenomeni di italianizzazione più evidenti nelle città).

Tabella 2 - *Valori LDN di oggetti linguistici parlati in Nord Italia e di controllo*

	LMO Ber.	LMO Pav.	PMS Tor.	EGL Bol.	RGN For.	VEC Ven.	ITA	ITA P. Rec.	SCN Pal.	FRA	CAT	JPN
LMO Bergamo	0.0000	0.4659	0.4556	0.4923	0.4874	0.5478	0.6024	0.5842	0.6754	0.7211	0.6201	0.9181
LMO Pavia	0.4659	0.0000	0.4701	0.5040	0.3838	0.4461	0.5471	0.5745	0.6661	0.7047	0.6272	0.9291
PMS Torino	0.4556	0.4701	0.0000	0.5201	0.5489	0.5203	0.5592	0.6092	0.6430	0.6926	0.6058	0.9218
EGL Bologna	0.4923	0.5040	0.5201	0.0000	0.4424	0.5060	0.5750	0.6008	0.6751	0.7453	0.6816	0.9114
RGN Forlì	0.4874	0.3838	0.5489	0.4424	0.0000	0.4839	0.5451	0.5455	0.6861	0.6670	0.6216	0.9245
VEC Venezia	0.5478	0.4461	0.5203	0.5060	0.4839	0.0000	0.3978	0.4694	0.6196	0.7333	0.5984	0.9139
ITA	0.6024	0.5471	0.5592	0.5750	0.5451	0.3978	0.0000	0.3804	0.4997	0.7273	0.5841	0.9025
ITA P. Recanati	0.5842	0.5745	0.6092	0.6008	0.5455	0.4694	0.3804	0.0000	0.5287	0.7356	0.6130	0.9228
SCN Palermo	0.6754	0.6661	0.6430	0.6751	0.6861	0.6196	0.4997	0.5287	0.0000	0.7933	0.6833	0.8815
FRA	0.7211	0.7047	0.6926	0.7453	0.6670	0.7333	0.7273	0.7356	0.7933	0.0000	0.7025	0.9579
CAT	0.6201	0.6272	0.6058	0.6816	0.6216	0.5984	0.5841	0.6130	0.6833	0.7025	0.0000	0.9459
JPN	0.9181	0.9291	0.9218	0.9114	0.9245	0.9139	0.9025	0.9228	0.8815	0.9579	0.9459	0.0000

Sulla scorta dei dati nella Tabella 2 possiamo fare alcune osservazioni. Per prima cosa, non vediamo mai il valore di *cut-off* di 0,51. Infatti, sebbene alcune coppie di oggetti vi si avvicinino (piemontese torinese ed emiliano bolognese, 0,5201; lombardo pavese ed emiliano bolognese, 0,5040), gli altri valori di LDN sono sempre piuttosto distanti dal valore soglia. Ciò era già stato notato da Wichmann (2019: 829) per i suoi dati, ed è bene ribadire l'importanza di questo fatto anche qui perché evidenzia ancora una volta come i casi più dubbi siano in realtà "atipici".

Venendo alla situazione del panorama linguistico del territorio nord-occidentale italiano, il cosiddetto "*continuum* gallo-italico" pare essere effettivamente un *continuum* che collega il piemontese al romagnolo passando per lombardo ed emiliano (PMS <> LMO <> EGL <> RGN), e questo proprio perché le singole coppie di oggetti linguistici adiacenti mostrano un livello di omogeneità linguistica inferiore a 0,51, anche se con un valore di LDN piuttosto vicino a quello di *cut-off* (caso raro interlinguisticamente). Ciò mostra come gli oggetti adiacenti siano varietà di una stessa lingua, ma con una diversità strutturale (lessico-fonetica) abbastanza alta. Infatti, i valori di LDN tra i singoli oggetti linguistici adiacenti sono: PMS <> LMO (0,46/0,47[8]), LMO <> EGL (0,49/0,50), EGL <> RGN (0,44). È da notare inoltre che anche per gli oggetti tradizionalmente considerati come varietà della stessa lingua LMO, come il bergamasco e il pavese, il valore di LDN è piuttosto vicino a 0,51 (0,47) e simile ai valori di LDN tra gli altri oggetti linguistici della zona.

I numeri che abbiamo appena presentato permettono anche qualche altra considerazione. Innanzitutto, EGL bolognese e RGN forlivese risultano essere piuttosto omogenei tra di loro, tanto che si potrebbe parlare per queste di due varietà di un'unica lingua (ovvero l'emiliano-romagnolo, v. anche oltre per i conteggi).

Esistono poi altre due situazioni interessanti che emergono osservando la Tabella 2 e che riguardano rispettivamente il veneto veneziano (VEC) e il siciliano palermitano (SCN). In relazione al piemontese, al lombardo, all'emiliano e al romagnolo, il veneto mostra un grado di omogeneità quasi sempre oltre il limite di 0,51 (LMO BG 0,55; PMS 0,52) o molto vicino ad esso (EGL 0,5060; RGN 0,48) con la sola ecce-

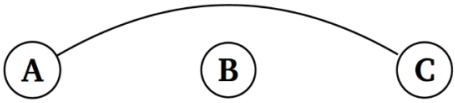
[8] Per semplicità, abbiamo riportato i valori di LDN approssimati a due cifre dopo la virgola seguendo la prassi di arrotondare per difetto se il valore era inferiore a 0,05 compreso e per eccesso se il valore era superiore a 0,051 compreso.

zione del lombardo pavese (0,45). Questa situazione posizionerebbe il veneto leggermente al di fuori del *continuum* gallo-italico, ma senza escluderlo totalmente: sarebbe collegato al *continuum* tramite il romagnolo e, seppur solo parzialmente, l'emiliano, ma non dal lato occidentale in cui abbiamo l'unica vera cesura con il lombardo bergamasco (anche se essa non si verifica con il lombardo pavese, per il quale però può influire una serie di fenomeni di italianizzazione che discuteremo brevemente sotto). Allo stesso tempo il veneto presenta un valore LDN decisamente inferiore rispetto all'italiano standard (0,40). Perciò il veneto sarebbe una lingua diversa da piemontese e lombardo bergamasco (ma non dal lombardo pavese e dall'emiliano-romagnolo), ma sarebbe anche da considerare come una varietà dell'italiano. Il siciliano, invece, risulta essere distante dagli oggetti linguistici propri delle regioni del Nord Italia (valori tra lo 0,64 e lo 0,68), come ci si aspetterebbe, ed è una lingua diversa anche dal portorecanatese (0,53), ma non dall'italiano standard (0,50), sebbene questi ultimi due oggetti siano da considerarsi come due varietà della stessa lingua (0,38). Queste due situazioni paiono piuttosto bizzarre, ma possono in realtà fungere da cartina al tornasole di alcuni problemi del metodo dell'ASJP, su cui diremo al §4.

Torniamo ora all'area di pertinenza di questo contributo. Abbiamo visto come il panorama linguistico gallo-italico sia effettivamente da considerarsi un *continuum*. La domanda che ci potremmo porre ora è se sia possibile contare il numero di lingue all'interno di un *continuum*, e, se è possibile contarle, quante lingue ci siano nel nostro. Sebbene la nozione di *continuum* sembri essere in contraddizione con il conteggio del numero di lingue in esso contenute (a causa della natura discreta del conteggio stesso), Hammarström (2008) propone un metodo che permette, anche in queste situazioni, di contare matematicamente il numero di lingue andando a identificare il numero minimo di 'confini' di omogeneità che esistono all'interno di una catena di oggetti linguistici. Secondo Hammarström (2008: 37), avendo un numero X di oggetti linguistici in un *continuum*, il numero minimo di lingue al suo interno sarà pari al numero di partizioni (*k*) minimo che possiamo identificare per raggruppare oggetti omogenei. Facciamo un breve esempio. Poniamo di avere un *continuum* linguistico Z composto da tre oggetti linguistici che chiameremo A, B e C. In Z, A e B sono omogenei e B e C lo sono altrettanto tra di loro, invece A e C

non sono omogenei. In Figura 1 riportiamo questa situazione rappresentata tramite una notazione che prevede una linea di congiunzione tra gli oggetti linguistici non omogenei.

Figura 1 - *Rappresentazione del continuum linguistico Z*

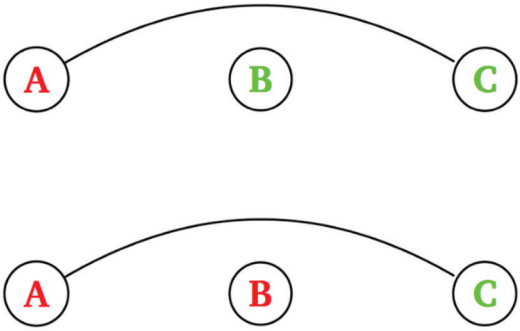


Da un punto di vista combinatorio, abbiamo diverse possibilità di dividere questo *continuum*: (a) tramite una sola partizione che possiamo rappresentare in questo modo Z = {A, B, C}[9]; (b) tre opzioni diverse che prevedono una partizione in due (Z = {A, B}, {C}; Z = {A}, {B, C}; Z = {A, C}, {B}); (c) una sola possibilità di partizione in tre, Z = {A}, {B}, {C}. Ovviamente, queste sono possibilità puramente teoriche e solamente tre di queste possono effettivamente avere un giudizio di verità positivo per il nostro *continuum* Z, ovvero due delle tre possibilità di partizione in due (Z = {A, B}, {C} e Z = {A}, {B, C}, ma non Z = {A, C}, {B}) e l'unica possibilità di partizione in tre (Z = {A}, {B}, {C}[10]). Pertanto, abbiamo due possibilità di valori per *k*, 2 e 3, ma il numero minimo di partizioni è 2. Quindi, nel nostro *continuum* Z possiamo contare un numero minimo di due lingue, come riportato a colori in Figura 2, e questo indipendentemente dal posizionamento dell'oggetto "intermedio" B, cioè non potremo decidere quale delle due partizioni in due gruppi sia da preferire.

[9] Gli oggetti linguistici racchiusi da una stessa coppia di parentesi graffe sono da considerarsi omogenei.

[10] Sebbene quest'ultimo caso appaia fuorviante, Hammarström (2008) ritiene che vada comunque considerato tra le possibilità di partizione perché la condizione di verità risiede nell'impossibilità di avere un "blocco" di oggetti linguistici in cui questi siano non omogenei e perciò la partizione in tre non si può escludere a priori.

Figura 2 - *Partizioni minime per il continuum Z*



La situazione degli oggetti linguistici gallo-italici da noi selezionati sembra essere un po' più complessa. Infatti, abbiamo il piemontese che risulta essere omogeneo con il lombardo (0,46/0,47), ma non con l'emiliano (0,52) e il romagnolo (0,55) (PMS: ≃ LMO / ≄ EGL, RGN); il lombardo che è omogeneo con il piemontese (0,46/0,47), con l'emiliano (0,49/0,50) e con il romagnolo (0,49/0,38) (LMO: ≃ PMS, EGL, RGN); l'emiliano che non è omogeneo con il piemontese (0,52), ma è omogeneo con il lombardo (0,49/0,50) e con il romagnolo (0,44) (EGL: ≃ LMO, RGN / ≄ PMS); e, infine, il romagnolo che non è omogeneo con il piemontese (0,55), ma è omogeneo con il lombardo (0,49/0,38) e con l'emiliano (0,44) (RGN: ≃ LMO, EGL / ≄ PMS). I valori di LDN per questi codici sono riassunti in Tabella 3.

Tabella 3 - *Valori di LDN per gli oggetti linguistici gallo-italici selezionati per il presente contributo*

		PMS	LMO		EGL	RGN
			BG	PV		
PMS		==	0,46	0,47	0,52	0,55
LMO	BG	0,46	==	0,47	0,49	0,49
	PV	0,47	0,47	==	0,50	0,38
EGL		0,52	0,49	0,50	==	0,44
RGN		0,55	0,49	0,38	0,44	==

Pertanto, sebbene la situazione gallo-italica appaia più complessa, è del tutto analoga a quella che abbiamo visto poco sopra per il *continuum* fittizio, come illustrato dalla Figura 3.

Figura 3 - *Rappresentazione delle partizioni minime per il* continuum *gallo-italico*

Abbiamo una situazione in cui gli unici valori di LDN superiori alla soglia di *cut-off* sono tra il piemontese e l'emiliano-romagnolo (segnalati con due colori differenti, rosso e verde rispettivamente), il lombardo è omogeneo con tutti gli altri oggetti e quindi potrebbe fungere da oggetto di transizione (segnalato in giallo proprio perché unione dei due colori primari rosso e verde). Possiamo quindi affermare che nel *continuum* gallo-italico sono presenti almeno due lingue e che molto probabilmente esse sono rappresentate dai due poli opposti, piemontese da una parte ed emiliano-romagnolo dall'altra, mentre il lombardo rappresenterebbe una varietà-ponte. È bene ancora aggiungere una considerazione osservando i risultati della nostra analisi: nell'Italia nord-occidentale l'ASJP riconosce due lingue diverse, mentre invece Ethnologue (che assegna ufficialmente i codici ISO 639-3) ne registra quattro, appunto PMS, LMO, EGL e RGN.

4. *Conclusioni e scenari futuri*

La differenza tra i due conteggi che abbiamo appena riscontrato non sorprende. Una delle conclusioni più robuste del lavoro di Wichmann, infatti, è la constatazione che "Ethnologue tends to overdifferentiate" (Wichmann 2019: 829) cosicché il numero dei codici ISO 639-3 assegnati è superiore all'effettivo numero di lingue diverse parlate nel mondo.

Nondimeno, dal nostro saggio di applicazione dell'ASJP alle varietà parlate in Italia emergono alcuni aspetti di questo metodo che è forse bene discutere.

In primo luogo, l'ASJP si configura sì come uno strumento automatico, come vuole anche il nome stesso del programma, quindi di

immediato utilizzo e di esito pressoché istantaneo ed esatto, almeno secondo i parametri prescelti, ma ciò aumenta la necessità di inserire nel programma dati, ovvero nel concreto liste di parole, che devono essere affidabili e accuratamente controllate da parte di chi compie la ricerca. Non si tratta di un rilievo marginale, dal momento che tra le liste che abbiamo impiegato per questa ricerca persino quella di una lingua con molti parlanti e molto studiata come l'italiano presentava alcuni errori (che noi abbiamo corretto, per questo lavoro). Alcuni erano evidenti refusi, come <dante> registrato in luogo di <dente> per l'entrata 'tooth' della lista; altri errori erano invece sviste più "sottili" e dovute spesso al fatto che le parole delle liste dell'ASJP sono trascritte con l'ASJPcode, una grafia *ad hoc* con la quale i compilatori possono anche avere poca familiarità.

C'è poi un altro potenziale problema (annotato già da Brown *et al.* 2013: 9), ovvero la differente natura delle liste redatte con l'ASJPcode: per lingue per le quali disponiamo di descrizioni minute e attendibili dell'inventario fonematico si opta generalmente per l'uso dell'ASJPcode partendo dalla trascrizione fonematica in IPA, ma la distanza strutturale potrebbe cambiare anche solo adottando la trascrizione fonetica. Altre liste – essendo il database a libero accesso – si basano talvolta sulle corrispondenze tra l'ASJPcode e le convenzioni grafiche (standard o no, riflesse o irriflesse, adoperate da esperti o da neofiti) dell'oggetto linguistico in questione. In queste condizioni, insomma, il rischio di inficiare, di tanto o di poco, il calcolo delle LDN è molto alto e lo stretto controllo, da parte di chi è competente in linguistica (o meglio nelle lingue dell'area interessata), diventa necessario.

Infine, visti alcuni risultati di questa nostra analisi, è senz'altro opportuno chiedersi se davvero sia sufficiente per il calcolo dell'omogeneità prendere in considerazione i soli dati lessico-fonetici. Come sappiamo, il lessico e la fonetica rappresentano la buccia delle lingue e sono, anche per varietà parlate in Italia, le parti del sistema più esposte al contatto e perciò al livellamento ad esso dovuto (Ricca 2010; Cerruti 2016). Altre indagini, sempre svolte guardando principalmente al lessico, ma questa volta coinvolgendo l'effettiva comprensione di questo all'interno di frasi reali attraverso il test SPIN (*Speech Perception In Noise*, v. Kalikov *et al.* 1977), hanno mostrato come il VEC, varietà triestina, non sia intercomprensibile con l'italiano standard (Tremul 2021): si tratta di un dato in forte contrasto con il no-

stro relativo al VEC. A questo punto, però, sorge la domanda relativa a quali strumenti adoperare per valutare le distanze inerenti al piano della morfologia o della sintassi: se non si ricorre ai test di riconoscimento, come appunto lo SPIN, è lecito chiedersi quali tratti morfosintattici occorrerebbe selezionare e quali trattare come più rilevanti per saggiare la distanza strutturale o la mutua intelligibilità tra oggetti linguistici[11].

In conclusione, il metodo dell'ASJP sicuramente fornisce un nuovo strumento metodologicamente solido per lo studio dell'omogeneità tra oggetti linguistici. Allo stesso tempo, però, nonostante i passi avanti, ancora molto lavoro è da fare per poterlo mettere a punto in modo che sia indiscutibilmente affidabile. A lavori futuri, nostri o altrui, rinviamo la riflessione su questi problemi e la proposta di adeguate risposte.

Ringraziamenti

Vogliamo ringraziare innanzitutto coloro che ci hanno fornito i loro dati, di prima mano, per gli oggetti linguistici discussi nel contributo: Salvo Baiamonte, Elia Calligari, Enrico Castro, Riccardo Mazzieri, Davide Tintori, Daniele Vitali. Ringraziamo, inoltre, i due revisori anonimi per i commenti e i suggerimenti puntuali che ci hanno permesso di migliorare la prima versione dell'articolo. Il lavoro è stato portato aventi e discusso congiuntamente dai due autori; per i soli fini accademici, Simone Mattiola è responsabile della stesura dei §§2 e 3, mentre Emanuele Miola della stesura dei §§1 e 4.

Riferimenti bibliografici

Ambrazas, Vytautas (a cura di). 1997. *Lithuanian grammar*. Vilnius: Baltos lankos.

Berruto, Gaetano. 1993. Le varietà del repertorio. In Sobrero, Alberto A. (a cura di), *Introduzione all'italiano contemporaneo. II. La variazione e gli usi*, 3–36. Roma-Bari: Laterza.

[11] Tra i metodi di questo tipo merita comunque una menzione il *Parametric Comparison Method* – PCM (v. Longobardi & Guardiano 2017), che si è provato ad applicare anche agli oggetti linguistici parlati in Italia.

Berruto, Gaetano. 2009. Lingue minoritarie. In Aa.Vv., *XXI Secolo. Comunicare e rappresentare*, 335–346. Roma: Istituto della Enciclopedia Italiana.

Brown, Cecil H. & Holman, Eric W. & Wichmann, Søren. 2013. Sound correspondences in the world's languages. *Language* 89(1). 4–29.

Cerruti, Massimo. 2016. L'italianizzazione dei dialetti. Una rassegna. *Quaderns d'Italià* 21. 63–74.

Cravens, Thomas D. 2022. Dialetto, dialect e l'incomprensione all'estero. In Giannelli, Luciano (a cura di), *Tra Po e Tevere, e altre terre e altri mari: studi di lingua e di culture*. 273–290. Bologna: Pendragon.

Cysow, Michael & Good, Jeff. 2013. Languoid, Doculect and Glossonym: Formalizing the Notion 'Language'. *Language Documentation & Conservation* 7. 331–359. (http://hdl.handle.net/10125/4606) (Consultato il 22.11.2024.)

Devoto, Giacomo. 1970. L'Italia dialettale. In Aa.Vv., *I dialetti dell'Italia mediana con particolare riguardo alla regione umbra*, 93–127. Gubbio: Centro Studi Umbri.

Francescato, Giuseppe. 1972. La classificazione delle parlate romanze: Alcuni problemi di metodo. *Romania* 5. 131–140.

Francescato, Giuseppe. 1980. A proposal for the classificazion of Romance Languages. In Izzo, Herbert J. (a cura di), *Italic and Romance*, 75–83. Amsterdam: John Benjamins.

Goebl, Hans. 1987. Points chauds de l'analyse dialectométrique: pondération et visualisation. *Revue de linguistique romane* 51. 63–118.

Goebl, Hans. 2008. La dialettometrizzazione integrale dell'AIS. Presentazione dei primi risultati. *Revue de linguistique romane* 72. 25–113.

Grandi, Nicola. 2018. Che tipo, l'italiano neostandard! In Moretti, Bruno & Kunz, Aline & Natale, Silvia & Krakenberger, Etna (a cura di), *Le tendenze dell'italiano contemporaneo rivisitate. Atti del LII Congresso Internazionale di Studi della Società di Linguistica Italiana (Berna, 6-8 settembre 2018)*, 59–74. Milano: Officinaventuno.

Grandi, Nicola. 2020. La diversità inevitabile. La variazione linguistica tra tipologia e sociolinguistica. *Italiano LinguaDue* 1. 416–429.

Grassi, Corrado & Sobrero, Alberto A. & Telmon, Tullio. 2001. *Introduzione alla dialettologia italiana*. Roma-Bari: Laterza.

Grimes, Joseph E. & Agard, Frederick B. 1959. Linguistic Divergence in Romance. *Language* 35(4). 598–604.

Hammarström, Harald. 2008. Counting languages in dialect continua using the criterion of mutual intelligibility. *Journal of Quantitative Linguistics* 15(1). 34–45.

Haspelmath, Martin. 2013. Argument indexing: A conceptual framework for the syntactic status of bound person forms. In Bakker, Dik & Haspelmath, Martin (a cura di), *Languages across boundaries*, 209–238. Berlin/New York: Mouton de Gruyter.

Haspelmath, Martin. 2021. Towards standardization of morphosyntactic terminology for general linguistics. In Alfieri, Luca & Arcodia, Giorgio Francesco & Ramat, Paolo (a cura di), *Linguistic categories, language description and Linguistic Typology*, 35–58. Amsterdam/Philadelphia: Benjamins.

Haspelmath, Martin. 2023. Types of clitics in the world's languages. *Linguistic Typology at the Crossroads* 3(2). 1–59. (https://typologyatcrossroads.unibo.it/article/view/16057) (Consultato il 22.11.2024.)

Iliescu, Maria. 1969. Ressemblances et dissemblances entre les langues romanes du point de vue de la morpho-syntaxe verbale. *Revue de linguistique romane* 33. 113–132.

Jäger, Gerhard. 2015. Support for linguistic microfamilies from weighted sequence alignment. *PNAS* 112(41). 12752-12757.

Kalikov, Daniel N. & Stevens, Kenneth N. & Elliott, Lois L. 1977. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *Journal of the Acoustic Society of America* 61(5). 1337–1351.

Longobardi, Giuseppe & Guardiano, Cristina. 2017. Phylogenetic reconstruction in syntax: the Parametric Comparison Method. In Ledgeway, Adam & Roberts, Ian (a cura di), *The Cambridge handbook of historical syntax*, 241–271. Cambridge: Cambridge University Press.

Mathiassen, Terje. 1996. *A short grammar of Lithuanian*. Columbus, OH: Slavica Publishers.

Miola, Emanuele. 2020. Che differenza c'è tra lingua e dialetto? https://www.linguisticamente.org/che-differenza-ce-tra-lingua-e-dialetto/ (Consultato il 22.11.2024.)

Muljačić, Žarko. 1967. Die Klassifikation der Romanischen Sprachen. *Romanistisches Jahrbuch* 18(1). 23–37.

Pei, Mario A. 1949. A new methodology for Romance classification. *Word* 5. 135–146.

Pellegrini, Giovan Battista. 1970. La classificazione delle lingue romanze e i dialetti italiani. *Forum Italicum* 4(2). 211–237.

Ricca, Davide. 2010. Italianizzazione dei dialetti. In Simone, Raffaele (a cura di), *Enciclopedia dell'italiano*, 711–713. Roma: Istituto dell'Enciclopedia Italiana Treccani.

Schmalstieg, William R. 1993. Le lingue baltiche. In Giacalone Ramat, Anna & Ramat, Paolo (a cura di), *Le lingue indoeuropee*, 481–506. Bologna: Il Mulino.

Simons, Gary F. & Fennig, Charles D. 2017. *Ethnologue: Languages of the world, twentieth edition*. Dallas, TX: SIL International.

Sobrero, Alberto A. 1993. *Dialetti in Italia*. Roma: Istituto della Enciclopedia Italiana Treccani.

Swadesh, Morris. 1950. Salish internal relationships. *International Journal of American Linguistics* 16(4). 157–164.

Tamburelli, Marco & Brasca, Lissander. 2018. Revisiting the classification of Gallo-Italic: a dialectometric approach. *Digital Scholarship in the Humanities* 33(2). 442–455.

Tremul, Teresa. 2021. Intercomprensione: il difficile rapporto tra dialetto triestino e italiano. Università di Bologna (Tesi di laurea triennale.)

Wichmann, Søren. 2019. How to distinguish languages and dialects. *Computational Linguistics* 45(4). 823–831.

Wichmann, Søren & Holman, Eric W. & Brown, Cecil H. (a cura di). 2022. *The ASJP Database* (version 20). https://asjp.clld.org/ (Consultato il 22.11.2024.)