

ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA

DIPARTIMENTO DI LINGUE LETTERATURE E CULTURE MODERNE

Corso di Studio in Lingua e cultura italiane per stranieri

**Lingue con e senza tradizione scritta:
costruzione di un campione tipologico**

PROVA FINALE IN DIVERSITÀ LINGUISTICA

Relatore:

Caterina Mauri

Presentata da:

Francesca Pinna

Correlatore:

Francesca Masini

Sessione marzo 2024

Anno accademico 2022 - 2023

Indice

Introduzione.....	1
1. Tipologia linguistica e campioni di lingue	4
1.1. La tipologia linguistica.....	4
1.2. Campioni tipologici: come vengono costruiti	6
1.2.1. Campioni di varietà e campioni probabilità	9
2. Lingue scritte e lingue orali.....	12
2.1. Universali del parlato	12
2.2. Caratteristiche della lingua scritta	18
2.3. Bias della lingua scritta	22
3. Materiali e metodi	26
3.1. Risorse utilizzate	26
3.1.1. World Atlas of Language Structures (WALS).....	26
3.1.2. Ethnologue: Languages of the World.....	27
3.1.3. Grambank	30
3.1.4. Glottolog.....	31
3.2. Procedura di costruzione dei campioni	31
3.2.1. Focus sulle criticità- Lingua Maung.....	34
3.3. Distribuzione dei fenomeni nei campioni	36
3.3.1 Creazione database	36
3.3.2. Analisi database.....	38
4. Risultati	43
5. Discussioni e conclusioni	54
Bibliografia.....	58
Sitografia	64
Appendice.....	65

Tabella S1: Campione lingue CON tradizione scritta	65
Tabella S2: Campione lingue SENZA tradizione scritta	72
Tabella S3: Campione WALS.....	80
Ringraziamenti	89

Introduzione

Secondo i dati riportati su *Ethnologue: Languages of the World*, fonte di riferimento di dati sulle lingue del nostro pianeta, il calcolo sul numero dei sistemi linguistici del mondo varia da 4.000 a 10.000. Nel 2022 si contavano 7151 lingue vive nel mondo. Ad eccezione di alcune lingue ricche di studi e ricerche, esistono molte di queste che non sono documentate, parlate da poche persone e in pericolo d'estinzione. 2,895 lingue sono minacciate o in via di estinzione e si prevede che tra il 50 e il 90 per cento si estingueranno nei prossimi 100 anni (Krauss 1992; Miyaoka et al. 2007; Austin e Sallabank 2011, Nettle 2000, Crystal 2002).

La distribuzione delle lingue nel mondo è diseguale. Solo il 4% delle lingue proviene dal continente europeo, ma queste sono parlate come lingua materna da più di un quarto della popolazione. La parte più considerevole della diversità linguistica del pianeta si trova, invece, nel continente americano e nella regione oceanica e del Pacifico, ma tantissime di queste lingue sono in via d'estinzione a causa dell'uso che ne viene fatto e del limitato numero di parlanti. Tra le lingue che rischiano di scomparire, molte di quelle che non sono state ancora adeguatamente documentate sono lingue non scritte (Adda et al., 2016). Circa la metà delle circa 7.000 lingue del mondo non hanno un'ortografia e quindi nessuna documentazione scritta (Lewis, 2014). Ad oggi, non si dispone di un metodo sistematico per raccogliere dati sul numero di comunità che stanno sviluppando le loro lingue, quale stadio hanno raggiunto, se i sistemi di scrittura esistenti vengono effettivamente utilizzati o se sono stati fatti tentativi per sviluppare sistemi di scrittura che non sono in uso. Inoltre, la maggior parte di queste lingue sono parlate da popolazioni relativamente piccole in Africa, Asia, Pacifico e America Latina, zone in cui, come anticipato, è presente una grande varietà linguistica (UNESCO, 2022). Queste minoranze linguistiche sono spesso gruppi di persone emarginate che vivono all'ombra delle lingue principali o sono isolati dai centri urbani dotati di maggiore ricchezza, opportunità educative e assistenza sanitaria. A meno che queste lingue non vengano documentate prima che cadano in disuso, molte informazioni linguistiche, culturali e antropologiche andranno perse per sempre. L'estinzione delle lingue non è soltanto una possibilità futura, ma è un fenomeno molto attuale: ogni giorno si perdono le competenze linguistiche di

coloro che parlano lingue prive di documenti scritti, contribuendo a un processo che ha visto molte lingue estinguersi nel corso del XX secolo. Ethnologue segnala 367 lingue che si sono estinte dal 1950 a un tasso medio di 6 lingue all'anno (Lewis, 2014).

La nostra concezione della lingua è profondamente influenzata da una lunga tradizione di analisi solo della lingua scritta (Linell, 1982). Ci sono migliaia di lingue a rischio di estinzione per le quali non esiste quasi nessuna documentazione e il tempo stringe prima che scompaiano (Adda et al., 2016). La lingua media la conoscenza dei gruppi degli esseri umani, la loro storia e l'importanza di alcuni concetti nelle diverse culture. Pertanto, senza documentazione, gran parte della conoscenza posseduta da chi parla una lingua non scritta muore insieme a questa. In molti casi, tuttavia, le lingue non muoiono ma subiscono un processo “*shift*” linguistico, ma alcune informazioni come i rituali tradizionali o la conoscenza del territorio non vengono trasmesse ai parlanti più giovani. In questi casi, gran parte della conoscenza muore anche quando la lingua continua ad essere parlata (Adams, 2017).

La presente tesi esplora le differenze intrinseche tra lingue scritte e lingue orali, mettendo in luce il fenomeno del bias della lingua scritta derivante dall'asimmetria tra queste due modalità espressive. L'approccio usato all'interno della tesi ha un duplice obiettivo. Il primo, metodologico, è quello di costruire due campioni di circa 200 lingue, facendo riferimento al campione *World Atlas of Language Structures* (WALS.info, Dryer & Haspelmath, 2013), in cui vengono raggruppate, da un lato, tutte le lingue orali, che non hanno una tradizione scritta e, dall'altro, quelle che invece, oltre ad essere orali, ce l'hanno. Attraverso un'analisi approfondita dei campioni si arriva al secondo obiettivo, che è quello di verificare, per un numero di interrogativi ristretto, cosa accade al variare di questo parametro, quindi di verificare questo *bias* nei fenomeni scelti.

Nel tentativo di esaminare in maniera approfondita il tema trattato, il seguente elaborato è composto da 5 capitoli. La prima parte è di tipo teorico e la seconda dedicata alla ricerca empirica, all'analisi dei dati e alle discussioni. Nello specifico:

- nel primo capitolo viene delineato un panorama della tipologia linguistica, fornendo definizioni dei concetti chiave relativi a questo campo e un contesto storico delle teorie e degli studi linguistici. Si tracciano le origini e lo sviluppo delle nozioni di tipologia linguistica nel corso del tempo, analizzando come queste abbiano influenzato l'approccio allo studio delle lingue. Vengono

esaminare le diverse categorie e classificazioni utilizzate per comprendere e categorizzare le lingue, con particolare attenzione alle differenze tra le lingue scritte e orali. La parte finale del capitolo si concentra sulla metodologia per la costruzione di campioni tipologici. Vengono esaminate le strategie di selezione dei dati che permettono di creare dei campioni che siano bilanciati.

- Il secondo capitolo si concentra su una revisione critica della letteratura linguistica, esaminando studi precedentemente condotti sulle caratteristiche della lingua scritta e della lingua orale, esplorando come la struttura, la sintassi, e la scelta lessicale possano variare in base ai due diversi tipi di comunicazione, fornendo esempi concreti di come queste differenze si manifestino. Si pone particolare attenzione sul bias della lingua scritta, prendendo in considerazione le cause che possono accentuare questo bias.
- Il terzo capitolo riguarda la parte operativa del lavoro di ricerca: sono presentate le risorse utilizzate per la creazione di due campioni tipologici, i materiali, i metodi e gli obiettivi facendo un particolare focus sulle criticità riscontrate durante l'elaborazione dei campioni e sulle relative soluzioni adottate per risolverle. È presente anche una descrizione del lavoro di estrazione semiautomatica svolto per l'analisi della distribuzione dei fenomeni scelti.
- Il quarto capitolo è dedicato ai risultati della tesi in cui sono emerse varie conclusioni significative derivanti dall'analisi sistematica dei dati raccolti. I risultati sono stati organizzati e resi più accessibili attraverso l'utilizzo di grafici e tabelle. I dati ottenuti sono stati analizzati in relazione agli obiettivi di ricerca.
- Nel quinto capitolo sono state discusse in dettaglio le implicazioni e le interpretazioni dei risultati emersi. Le conclusioni sono state contestualizzate rispetto al quadro teorico di riferimento, evidenziando collegamenti e discrepanze. Si è approfondito il confronto con la letteratura esistente, identificando punti di convergenza e divergenza. Inoltre, sono state proposte eventuali raccomandazioni per ricerche future o applicazioni pratiche, sottolineando l'importanza delle discussioni nel contesto della ricerca condotta.

1. Tipologia linguistica e campioni di lingue

1.1. La tipologia linguistica

La tipologia linguistica è una branca della linguistica che cerca di comprendere la natura del linguaggio umano analizzando la variazione interlinguistica. Il termine tipologia, in linguistica, rimanda alla classificazione delle lingue in tipi strutturali. Ciononostante, l'obiettivo finale della tipologia non è la classificazione in sé, ma è quello di trovare, all'interno delle lingue del mondo, qualcosa di comune, altamente frequente e prevedibile (Grandi & Mauri, 2022).

All'inizio dell'Ottocento, la scoperta di W. von Humboldt della diversità linguistica, grazie al contatto con le lingue non europee che non si adattavano al modello di grammatica generale a base greco-latina, ha portato all'inizio degli studi tipologici. Successivamente, Friedrich Schlegel (1808) e suo fratello August Wilhelm (1818) hanno proposto la prima classificazione tipologica. La tipologia classica di August Wilhelm von Schlegel e Wilhelm von Humboldt portata avanti nel Novecento da Sapir (1921), aveva come scopo quello di offrire una caratterizzazione globale dell'intera lingua basandosi su un numero ristretto di caratteristiche tipologiche. La morfologia, ovvero il modo in cui le lingue costruiscono le parole, era al centro di questi studi. La classificazione più famosa è basata sulle definizioni di A. Schleicher (1848) e, in particolare, di A. F. Pott (1849), che distingue quattro tipi linguistici: isolante, caratterizzato da lingue con poca o nulla morfologia, prevalenza di parole monosillabiche e dall'assenza di affissi (per es. il cinese); agglutinante, caratterizzato da lingue con morfologia articolata ma con morfemi facilmente separabili che hanno significati univoci e prevedibili, e da affissi a funzione univoca e basso grado di fusione con la radice (per es. il turco); flessivo o fusivo, caratterizzato da un cumulo di più significati sui singoli morfemi, da affissi a funzione multipla e fusi con la radice (per es. l'italiano); incorporante o polisintetico, in cui molti elementi della frase si fondono in parole estremamente lunghe e complesse e il confine tra parola e frase è difficile da tracciare (per es. l'eskimo) (Treccani IV Appendice, 1981).

A partire dalla metà del Novecento, il campo della tipologia linguistica moderna ha cominciato a prendere forma, principalmente grazie alle teorie introdotte da Joseph Greenberg. Questo nuovo approccio si focalizza sull'analisi di fenomeni e strutture grammaticali specifiche anziché sulla mera classificazione delle lingue nel loro complesso. Le teorie strutturaliste costituiscono il fondamento di questa evoluzione nella tipologia linguistica moderna, evidenziando un ruolo secondario della morfologia che agisce indipendentemente dai legami di parentela tra le lingue.

Dopo un rinnovato interesse per il principio della relatività linguistica proposto da Edward Sapir e Benjamin Lee Whorf, la ricerca si concentra sugli universali linguistici, ovvero sulle proprietà comuni a tutte le lingue del mondo. Si è cercato di riformulare i principi della tipologia linguistica a partire dalla constatazione che i tipi linguistici tradizionali non sono soddisfacenti. L'approccio contemporaneo non si limita a registrare e raggruppare le somiglianze, poiché ciascuna lingua realizza in sé più di un tipo e quindi il "tipo" è un costrutto teorico ideale (V. Skalička, 2013). La nozione di tipo fa riferimento ai modi in cui le lingue si comportano rispetto a un dato *tertium comparationis*. Secondo Seiler il *tertium comparationis* deve essere “*an integrated part of language activity, but not identical with an individual language*” (Seiler, 2000). Quindi, è fondamentale definire un *tertium comparationis* in ogni processo comparativo che coinvolga due o più lingue, ovvero un elemento terzo che resta sempre invariabile e costante. Infatti, si può osservare diversità quando si è in grado di fare un confronto tra due o più fatti linguistici (comparanda) e questo implica l'identificazione di qualche aspetto invariante che permette il confronto, cioè il *tertium comparationis* (Croft, 2002). Maggiore è il grado con cui il *tertium comparationis* individuato si manifesta, maggiore è l'ampiezza dell'interrogativo di ricerca. Una volta individuato il *tertium comparationis* si possono raggruppare i comportamenti attestati nelle diverse lingue, definendo diversi tipi di diversità linguistica: fonetica; lessicale, morfosintattica; pragmatica.

1.2. Campioni tipologici: come vengono costruiti

Ultimamente, con l'avvento dell'era digitale, con lo sviluppo dei corpora e con l'urgenza della descrizione di lingue in via d'estinzione, il ruolo del dato diventa centrale e la linguistica si caratterizza sempre di più come scienza empirica. Come tutte le scienze empiriche, la tipologia linguistica basa le sue generalizzazioni sulla solidità del campione di dati che viene analizzato. Innanzitutto, il tipologo individua una prospettiva di analisi scegliendo un fenomeno, un parametro o un dominio rispetto al quale confrontare tra loro le lingue. Di conseguenza, una volta identificato l'interrogativo di partenza e definito il *tertium comparationis* per mezzo di uno o più concetti comparativi, è necessario decidere quante e quali lingue servono per poter rispondere alla domanda di ricerca, nonché quanti e quali dati vanno raccolti (Grandi & Mauri, 2022).

I dati linguistici vengono distinti da Lehmann tra *primary data* e *secondary data*. I *primary data* sono rappresentazioni di specifici eventi discorsivi insieme alle loro coordinate spazio-temporali, cioè oggetti con un'identità storica. I *secondary data* sono rappresentazioni più astratte, con gradi diversi di astrazione, per esempio frasi estrapolate dal contesto, usate come *types* invece che come *tokens* oppure giudizi metalinguistici e valutazioni grammaticali (Lehmann 2004: 11). I dati linguistici possono essere:

- già esistenti, già raccolti e già organizzati (es. Corpus Repubblica, BNC, etc.)
- già esistenti, già raccolti e non ancora organizzati (dati linguistici sul web, insieme dei testi di una biblioteca digitale, etc.)
- già esistenti, non ancora raccolti e non ancora organizzati (archivi di trasmissioni televisive da trascrivere, selezione di testi scritti in un particolare periodo, etc.)
- non ancora esistenti, non ancora raccolti e non ancora organizzati: i dati vanno elicitati, cioè occorre stimolare la produzione (spontanea o controllata) di dati da raccogliere per la nostra ricerca

I dati possono essere ottenuti tramite diverse tecniche. (Lehmann 2004, Iannaccaro 2000, Turchetta 2000). Una di queste è l'introspezione (Iannaccaro 2000). In questo caso il linguista che conosce la lingua produce enunciati e li interpreta (giudizi di grammaticalità, parafrasi, etc.). Questo è il metodo tipico degli approcci più 'logici' della

linguistica e ed è problematico perché è poco oggettivo, circolare e produce idioletti. In alternativa, i dati possono essere “prodotti” o “trovati” (Iannaccaro 2000), il che significa che il linguista può partecipare attivamente alla produzione dei dati oppure può trovare dati preesistenti (pre-prodotti indipendentemente da lui). Quindi, può interpretare enunciati prodotti da altri, partendo dalla forma e dalla struttura (di solito da un corpus), fino ad arrivare alla funzione. Questo approccio è detto semasiologico. Per quanto riguarda l’approccio onomasiologico, invece, gli enunciati sono prodotti sotto il controllo del linguista, a partire dalle funzioni, attraverso un processo produttivo, si arriva alle forme e strutture.

In ogni ricerca tipologica la costruzione di un campione di lingue costituisce, quindi, una fase fondamentale. Domande di ricerca diverse richiedono diversi tipi di campioni; pertanto, l'obiettivo deve essere quello di costruire il campione di lingue più adatto per rispondere all'interrogativo di ricerca individuato. Nell'ambito dello studio delle 7.139 lingue parlate nel mondo, sarebbe auspicabile raccogliere dati da tutte queste lingue per fare affermazioni generali valide. Tuttavia, ci sono molti motivi per cui questo non è fattibile. Uno di questi è la diversità socioculturale e geografico-politica delle comunità che le parlano. Inoltre, molte lingue non sono state sufficientemente documentate, limitando un'analisi dettagliata di fenomeni specifici. Ci sono ancora numerose lingue che, sebbene esistano da diversi decenni, non sono state descritte in modo completo e sistematico, e altrettante sono le lingue ancora in attesa di qualche specialista che le descriva. La maggior parte delle lingue parlate in piccole comunità non ha mai sviluppato un sistema di scrittura o una tradizione grammaticale, al contrario di lingue, come l'italiano o il cinese, che vantano una lunga tradizione scritta, letteraria e studi linguistici approfonditi. Il semplice fatto che una lingua sia ufficialmente riconosciuta e inserita nell'attuale lista di Ethnologue non garantisce la presenza di una descrizione grammaticale esaustiva per quella specifica lingua. Inoltre, le fonti potrebbero non essere accessibili al tipologo, anche se esistono, perché magari si trovano in luoghi remoti o non sono disponibili per la consultazione. Infine, anche ammesso l'accesso a fonti di dati adeguate per tutte le 7.139 lingue, includerle tutte sarebbe (per la maggior parte degli interrogativi di ricerca) eccessivamente dispendioso in termini di tempo includerle tutte. Di conseguenza, diventa cruciale sviluppare un campione più limitato di lingue rappresentative, in grado di riflettere la varietà linguistica esistente e di generare

risultati universali che possano estendersi anche alle lingue non incluse nel campione. (Grandi & Mauri, 2022).

Recenti studi quantitativi sulla diversità linguistica mostrano che la diversità linguistica è influenzata e determinata da diversi fattori (Nichols 1992; Nettle 1999; Holman et al. 2007; Wichmann 2008, 2011), in particolare: vicinanza geografica, vicinanza genealogica mutamento linguistico. Per quanto riguarda ciò che Wichmann (2008) chiama *language dynamics*, dati empirici di variazione interlinguistica, vengono creati utilizzando simulazioni computazionali e algoritmi matematici. Questi consentono di distinguere e quantificare il ruolo che diversi parametri svolgono nello sviluppo della diversità linguistica (Holman et al. 2007; Georgi et al. 2010). Possibili distorsioni genealogiche, areali, tipologiche e sociali possono influenzare i risultati di un'analisi tipologica, ovvero:

- Distorsioni dovute alla parentela genealogica: un potenziale sbilanciamento dei dati raccolti si manifesta quando alcune famiglie linguistiche o specifici rami all'interno di queste famiglie risultano essere sovra-rappresentati all'interno del campione selezionato.
- Distorsioni dovute alla vicinanza areale: le somiglianze linguistiche tra lingue diverse possono essere il prodotto di interazioni sociali e commerciali prolungate tra le comunità di parlanti, specialmente quando queste comunità sono geograficamente vicine. I fenomeni di convergenza e interferenza linguistica, derivanti da questo contatto, possono influenzare notevolmente il livello di variazione linguistica osservabile. Holman et al. (2007) sostengono che la diversità linguistica è minima tra lingue distanti meno di 1000 km, quindi aumenta con l'aumentare della distanza. Questo vale nel Nuovo Mondo, in Africa e in Eurasia, confermando gli effetti dell'Isolamento per Distanza (IBD), termine introdotto da Wright (1943) nel campo della popolazione genetica, riferendosi a situazioni in cui “*large genetic correlations are found among spatially proximal populations and drop off smoothly as the distances among the populations decrease*” (Holman et al. 2007).
- Distorsioni dovute alla somiglianza tipologica: è fondamentale tenere sotto controllo anche il grado di somiglianza tipologica tra le lingue prese

in considerazione, ad esempio focalizzandosi sul tipo di struttura morfologica o sull'ordine basico dei costituenti delle lingue esaminate.

- Distorsioni dovute alla organizzazione sociale e alla tradizione culturale delle comunità: la presenza o meno di una tradizione scritta, letteraria o culturale permette che alcune lingue siano più accessibili di altre. Un certo sbilanciamento verso queste lingue è inevitabile, quindi ogni campione sarà influenzato dalla possibilità per il tipologo di reperire informazioni su una data lingua. Ma è importante ridurre questo sbilanciamento al minimo.

È pertanto cruciale considerare attentamente tali fattori durante la costruzione del campione al fine di prevenire qualsiasi forma di distorsione. A tal proposito, il metodo "Genus-Macroarea", sviluppato da Miestamo, Bakker e Arppe (2016) offre un approccio per ottenere un campione di varietà rappresentativo e flessibile, stratificato e strutturato in modo da soddisfare le esigenze e le fasi di ricerca diverse. Il metodo include infatti diversi campioni, o livelli di campionamento (Miestamo et al., 2016, pp. 247-56): il campione genealogico (Genus Sample, GS), il campione di base (Core Sample, CS), il campione ristretto (Restricted Sample), il campione esteso (Extended Sample). Il campione genealogico si basa sul livello del *genus* (Dryer, 1989), ovvero il "ramo" all'interno delle famiglie linguistiche. Secondo Miestamo, per ottenere il campione genealogico, è necessario innanzitutto selezionare in modo casuale una lingua per ogni *genus* attestato. Successivamente, bisognerà verificare siano disponibili e accessibili in tutte le lingue del campione informazioni dettagliate e affidabili. Se per un dato *genus* non è possibile trovare una lingua con informazioni sufficienti per il fenomeno preso in esame, l'intero *genus* andrà eliminato dal campione, formando in questo modo il campione di base. Queste modifiche, però, possono portare a uno squilibrio geografico-areale del campione; quindi, le aree descritte meglio saranno più rappresentate. Occorrerà allora identificare un sottocampione all'interno del campione di base bilanciato per macroaree geografiche. Questo è chiamato campione ristretto e permette di avere una serie di dati affidabile anche come campione di probabilità. Inoltre, il campione di base può essere ampliato aggiungendo una o più lingue di uno o più rami, creando in questo modo il campione esteso.

1.2.1. Campioni di varietà e campioni probabilità

La qualità dei campioni linguistici nel lavoro tipologico è molto importante. Come è stato già detto in precedenza, i campioni possono subire vari tipi di bias se non viene utilizzata

una strategia di campionamento appropriata. Rijkhoff *et al.* (1993) propongono un metodo di campionamento in cui il fattore genetico sta alla base. La loro proposta si concentra fortemente sulle idee avanzate da Bell (1978), ma vuole fornire un modo più obiettivo per evitare bias genetici. Grazie alla loro tecnica, i campioni che verranno creati mostreranno al meglio la varietà delle lingue che esistono in tutto il mondo. Rijkhoff *et al.* (1993, p. 171) fanno una distinzione tra due approcci al campionamento linguistico, il campione di probabilità e il campione di varietà. Quale di questi sia il più adatto dipende in gran parte dal tipo di domanda a cui si cerca di rispondere.

I campioni linguistici di probabilità sono utilizzati per la verifica statistica delle tendenze e delle correlazioni. Se si è interessati principalmente a questo, allora le lingue del campione devono essere indipendenti, non dovrebbero essere correlate in termini di affiliazione genetica, distribuzione geografica, ecc. Questo perché solo nel caso di lingue indipendenti si possono fare generalizzazioni statisticamente valide. Questi campioni indicheranno la frequenza con cui si può osservare un dato fenomeno. (Es. se l'ordine SVO è più frequente o meno degli altri). Rivelano non solo se un fenomeno è attestato o meno, ma quanto è frequente (Grandi & Mauri, 2022). Perkins (1980) ha ideato un modo per creare campioni di probabilità e ha reso possibile utilizzare test statistici per scoprire se esistono correlazioni tra lingue (Perkins 1989 e Dryer 1989). Come precedentemente menzionato, l'approccio migliore per evitare distorsioni genetiche implica l'inclusione di lingue provenienti da *phyla* diversi. Tuttavia, questa prassi può condurre a un campione contenente meno di trenta lingue (secondo la classificazione di Ruhlen, 1987), che comprende anche le lingue isolate, non correlate ad altre lingue o famiglie linguistiche, ognuna delle quali può essere considerata un *phylum* separato. In molti casi, tuttavia, le indagini interlinguistiche richiedono campioni più estesi per raggiungere obiettivi più specifici, oltre a risultati più globali. Perciò, è chiaro che progettare campioni di probabilità rappresentativi in cui le lingue non siano geneticamente correlate è molto difficile.

I campioni linguistici di varietà, invece, mostrano quanta più varietà possibile nella realizzazione linguistica dei fenomeni presi in esame e fanno emergere anche le strategie o i tipi di espressione più rari nel dominio esplorato. Questo tipo di campione è particolarmente adatto per ottenere generalizzazioni universali (Grandi & Mauri, 2022). Secondo Rijkhoff *et al.* un campione di varietà non solo deve contenere almeno una lingua per ciascun *phylum*, ma il numero di lingue in un campione che appartengono allo stesso

phylum dovrebbe essere proporzionale alla diversità linguistica in quel particolare *phylum*. A prima vista ci si potrebbe aspettare che più lingue contiene un *phylum*, maggiore è la possibilità di trovare forme e strutture diverse. Tuttavia, le cose non sono così semplici. Alcune famiglie più piccole (ad esempio quella afroasiatica — 258 esemplari) sono più diversificate rispetto ai gruppi numericamente più grandi (come la famiglia Bantu ampia — 500 esemplari, o la famiglia oceanica — 426 esemplari) (Bell 1978: 146). Ciò suggerisce che non si dovrebbe fare troppo affidamento sul numero di lingue geneticamente correlate nel tentativo di creare la massima diversità linguistica nel campione. Pertanto, Rijkhoff *et al* non prendono in considerazione solo il numero delle lingue, ma piuttosto il modo in cui queste lingue sono storicamente collegate. In un campione di varietà (al contrario di quello di probabilità) è molto importante avere casi rari, poiché "*exceptional types test the theory*" (Perkins 1988: 367). Una teoria universale della grammatica deve prevedere le grammatiche di tutte le lingue, indipendentemente dalla loro origine genetica, tipo linguistico o posizione geografica, se vuole essere universalmente valida. Pertanto, è necessario esaminare il più possibile l'intera gamma di forme e costruzioni così come si presentano nelle lingue naturali, affinché una teoria sia tipologicamente e descrittivamente adeguata. In questo modo, si può essere certi che la teoria copra tutte le varianti e che non ci siano controesempi alle regole e ai principi della stessa teoria.

2. Lingue scritte e lingue orali

I diversi studi sulla lingua parlata e la lingua scritta concordano sul fatto che queste differiscano sistematicamente. Studiosi come Walter J. Ong, nel suo libro "*Orality and Literacy: The Technologizing of the Word*" del 1985, delinea in modo approfondito le differenze fondamentali tra le lingue orali e scritte, analizzando come tali differenze influenzino la percezione, la trasmissione del sapere e la struttura del pensiero umano. Secondo Miriam Voghera la differenza tra il lingua parlata e lingua scritta non si basa su variazioni strutturali, come le coniugazioni verbali o le regole di accordo, ma piuttosto sulle caratteristiche che riflettono le condizioni semiotiche e cognitive in cui si svolge naturalmente la parola. Questo implica che non è corretto parlare di grammatica della lingua parlata contrapposta, ad esempio, a quella scritta, ma di usi linguistici adeguati alle condizioni generali del discorso (Voghera 2013). Alcuni altri accademici, tra cui Auer e Linell, hanno dedicato le proprie ricerche all'analisi dettagliata delle peculiarità distintive tra lingue orali e scritte. Il loro lavoro ha messo in luce come questi aspetti rivestano un ruolo centrale nella dinamica comunicativa. La loro ricerca non solo offre una visione dettagliata delle differenze formali tra queste modalità linguistiche, ma si estende anche all'indagine delle implicazioni culturali, sociali e cognitive associate a tali variazioni.

2.1. Universali del parlato

La comunicazione parlata si acquisisce durante la socializzazione primaria dell'individuo, che inizia e si sviluppa per lo più nella prima infanzia come un elemento integrante e naturale delle consuete attività all'interno del contesto culturale quotidiano (Linell, 1982). L'acquisizione della lingua madre avviene attraverso dialoghi parlati e gran parte dell'esperienza linguistica è costituita da dialoghi durante tutta vita. Dall'inizio della loro vita, gli adulti interagiscono con i bambini in modo spontaneo e naturale. Non solo parlano ai bambini come se potessero capire, ma interpretano le vocalizzazioni, i gesti e i movimenti dei bambini come segni di intenzionalità, considerandoli come risposte comunicative (Voghera, 2017). Di conseguenza, il dialogo si configura come l'ambito naturale e predominante per l'acquisizione linguistica, costituendo il principale modello dei vari usi

della lingua (Voghera, 2013) e la comunicazione linguistica è primariamente orale sia dal punto di vista funzionale sia quantitativo, cfr. De Mauro (1971).

Un aspetto peculiare della lingua parlata consiste nella sua natura effimera, dinamica e mutevole. Inoltre, quando parla, l'oratore cerca di esercitare un'influenza sull'ascoltatore o su un gruppo di ascoltatori con l'intento di suscitare sensazioni, comprensione, emozioni o azioni specifiche. È da sottolineare che il parlante si affida non solo alle parole pronunciate, ma anche ai segnali non verbali e alla vasta gamma di conoscenze condivise tra lui e l'ascoltatore. Inoltre, tiene conto delle reazioni dell'ascoltatore e di altre caratteristiche del contesto circostante in cui si svolge l'atto comunicativo (Linell, 1982). Nel contesto di una conversazione naturale, si manifesta una fluida alternanza di scambi comunicativi tra i partecipanti, che richiede un continuo processo di produzione e comprensione delle informazioni in tempo reale (Voghera, 2013). Peter Auer (2009), sostiene, infatti, che una delle differenze fondamentali tra lingua parlata e lingua scritta risiede nella "linearità" temporale della lingua parlata. Ciò non implica che la scrittura non sia un processo che richieda tempo, bensì sottolinea che la temporalità caratteristica della comunicazione parlata deriva da un processo dinamico e interattivo tra parlante e ascoltatore (Linell, 1982).

La temporalità del parlato può essere vista, secondo Auer, da almeno tre prospettive differenti: la transitorietà, l'irreversibilità e la sincronizzazione.

- La transitorietà della lingua parlata rimanda a quello che Hockett (1960) chiama "*rapid fading*". Il concetto di "rapid fading" si riferisce alla natura effimera e transitoria del linguaggio parlato. Indica che il segnale acustico prodotto durante la comunicazione verbale ha la caratteristica di svanire rapidamente dopo essere stato emesso. Questa proprietà denota che una volta pronunciata, un'unità linguistica (come una parola o una frase) non persiste nel tempo, a differenza di altri segnali o modalità di comunicazione che possono essere osservati o conservati nel tempo in modi diversi (ad esempio, segni grafici o oggetti fisici). Il "presente" della lingua parlata è limitato all'arco temporale entro il quale chi parla e chi ascolta può conservarlo nella memoria (Auer, 2009). Il dialogo normalmente implica un basso grado di pianificazione linguistica da parte del parlante e un basso grado di attenzione selettiva da parte del ricevente. Questi sono inversamente proporzionali al ritmo di alternanza dei turni di conversazione

(Voghera, 2013). È importante sottolineare che ricordare ciò che è stato detto rappresenta un processo dispendioso per la nostra mente, suscettibile a errori e disaccordi. La transitorietà del parlato, quindi, incide notevolmente sulla responsabilità sia del parlante per quanto detto, sia dell'ascoltatore per quanto udito. Al contrario, raramente si verificheranno controversie su quanto scritto, a condizione che entrambe le parti abbiano accesso a tali scritti. Va considerato anche che il nostro processo mentale, riguardo alla produzione e all'elaborazione del linguaggio parlato, presenta degli altri limiti. È frequente che sia chi parla che chi ascolta si trovino in situazioni in cui possono “perdere il filo” quando si tratta di costruzioni complesse. Di conseguenza, nella comunicazione verbale si evitano spesso strutture linguistiche che richiedono un'elaborazione prolungata nel tempo. La scrittura, invece, è un atto privato ed è solo il prodotto finito di questo atto ad entrare in contatto con il destinatario (Auer, 2009).

- L'irreversibilità del parlato, secondo Auer (2009), si riferisce al concetto secondo cui ciò che è stato detto durante un'interazione verbale non può essere facilmente annullato o riparato. Una volta pronunciato, il linguaggio parlato non può essere cancellato o corretto nella stessa misura in cui si può fare con la scrittura. Questo concetto sottolinea ancora una volta la natura irreversibile ed effimera del linguaggio parlato, dove le parole e le frasi pronunciate possono avere un impatto immediato e non possono essere facilmente ritratte o modificate. Naturalmente, questo non indica che non si possa provare a rivedere e riformulare ciò che è stato detto. Tuttavia, questo processo richiede la volontà e la cooperazione dell'interlocutore. Di solito, i co-conversatori sono disposti a consentire riparazioni per lapsus innocui, ambiguità referenziale, espressioni vaghe ecc., ma saranno meno disposti in caso di segreti svelati "accidentalmente", insulti, ecc. La lingua scritta è ovviamente del tutto irredimibile non appena viene ricevuta, ma fino ad allora sarà trascorso molto tempo. Durante il processo di scrittura, tuttavia, lo scrittore sarà libero di rivedere quanto scritto (Auer, 2009). La caratteristica distintiva della comunicazione verbale risiede nella sua assenza di un supporto di memoria esterna, come potrebbe essere una pagina scritta. Le implicazioni derivanti dall'irreversibilità del linguaggio parlato sono chiaramente visibili durante i fenomeni noti come editing, che includono pause, correzioni spontanee e ripetizioni, determinando una frammentazione inevitabile della

comunicazione verbale e, di conseguenza, richiedendo strategie di riparazione. Questi fenomeni, presenti nella comunicazione orale, non esistono oppure vengono cancellati nei testi scritti (Auer,2009).

- Un terzo punto di vista sulla temporalità della lingua parlata coinvolge in modo significativo la relazione tra parlante e ascoltatore e l'immediatezza della situazione faccia a faccia. La sincronizzazione del parlato si riferisce al fenomeno secondo cui i partecipanti a una conversazione si adeguano e coordinano istantaneamente il loro linguaggio, sia a livello verbale che non verbale, durante l'interazione comunicativa. Questa sincronizzazione avviene in tempo reale, implicando un adattamento reciproco e spontaneo nei ritmi, nei toni, nelle pause e in altri aspetti della comunicazione verbale e non verbale. Questo processo sincronico è cruciale per la fluidità e l'efficacia della conversazione. Nella comunicazione orale, come in nessun'altra forma di interazione, il flusso di coscienza tra l'Io e l'altro è sincronizzato (Auer et al., 1999). Grazie a questa sincronizzazione, le costruzioni sintattiche del parlante vengono elaborate dal ricevente solo con un breve ritardo. Al contrario, la produzione e la ricezione nella lingua scritta sono separate da una notevole distanza di tempo. Inoltre, per la maggior parte delle persone la scrittura e la lettura avvengono a velocità estremamente divergenti, mentre il parlare e il sentire hanno un ritmo relativamente simile. Gli effetti sulla sintassi di questa sincronizzazione tra parlante e ascoltatore nella comunicazione faccia a faccia sono stati riconosciuti attraverso l'importante ricerca di Goodwin (1979, 1980, 1995). Secondo Goodwin esiste un coordinamento dialogico tra le attività di parlante e ricevente, che porta alla costruzione collaborativa di frasi da parte di più partecipanti a una conversazione. Anche Lerner (1996, 2004) ha analizzato le cosiddette "*collaborative turn sequences*" nelle conversazioni quotidiane che evidenziano la reciproca partecipazione e la sincronizzazione tra i partecipanti durante l'interazione. Inoltre, secondo Voghera la lingua parlata fa un uso più ampio del contesto non verbale rispetto alla comunicazione scritta. Pertanto, diventa cruciale per gli interlocutori evidenziare le parti essenziali dell'enunciato attraverso specifici elementi linguistici o prosodici durante l'interazione comunicativa. Ciò è principalmente dovuto al fatto che i parlanti condividono la stessa situazione comunicativa e possono fare riferimento ad essa senza la necessità di utilizzare

esplicitamente le parole per nominare gli elementi contestuali. Questa possibilità di riferimento diretto agli elementi situazionali ottimizza i tempi dei processi di produzione e ricezione linguistica. Per giunta, i parlanti non si affidano solo agli elementi della situazione comunicativa, ma utilizzano anche altre modalità comunicative: mimica facciale, gesti, prossemica (Voghera, 2013).. L'uso di questi elementi produce una specie di commento della sequenza verbale contribuendo ad arricchire il significato complessivo di un enunciato (Mana et al. 2006).

Il discorso parlato rappresenta, secondo Voghera (2013) un prodotto multidimensionale, in quanto sia il parlante che l'ascoltatore elaborano significati provenienti da molteplici fonti. Il discorso parlato è caratterizzato dalle strutture linguistiche che rispondono a requisiti contestuali ed elaborativi normalmente associati al parlato e all'ascolto. Ciò si traduce in alcune costanti linguistiche generali:

- **Ridondanza:** al fine di ridurre la possibilità di interruzioni che potrebbero compromettere la comunicazione, spesso si tende a produrre testi ridondanti per evitare la perdita di informazioni e per agevolare l'ascoltatore nell'interpretazione del messaggio (Biber, 1995). La ridondanza si manifesta non solo a livello delle strutture linguistiche, ma anche a livello del contenuto tematico. Il testo parlato non è caratterizzato da una progressione tematica lineare, ma spesso si sviluppa attraverso un parziale ritorno su ciò che è già stato menzionato e consiste quindi in un processo continuo di aggiunta di nuove informazioni, parafrasi, riformulazioni e di ripetizione di vecchie informazioni. Questo fenomeno è comune in tutti i testi parlati, ma naturalmente è più evidente nei testi argomentativi rispetto a quelli narrativi, come sottolineato da Sornicola nel 1981. La ridondanza tematica può essere prodotta da varie strategie testuali. Tra queste la più comune è certamente la parafrasi (Sornicola 1981), che consiste in un progressivo ampliamento del tema attraverso un processo di riformulazione.
- **Disfluenza:** la necessità di organizzare il discorso in tempo reale porta a frequenti interruzioni del flusso verbale, comunemente denominate disfluenze. In un contesto più ampio, queste rappresentano momenti di pausa essenziali per il parlante, fungendo come meccanismo di controllo nella pianificazione del discorso. Tali interruzioni diventano più frequenti all'aumentare della durata dell'enunciazione. I fenomeni disfluenti, di vario genere, possono influire sulla

sequenza verbale in modi diversificati, e si distinguono in disfluenze fonetiche e disfluenze testuali.

- Ripetizione. Le lingue orali, essendo legate all'interazione faccia a faccia, impiegano la memoria sonora e la ripetizione (Farrel, 1985). Le ripetizioni hanno origine fondamentalmente dall'elaborazione produttiva e ricettiva on-line, che richiede di ottimizzare la durata dell'elaborazione delle informazioni. Secondo Voghera (1992) ci sono diverse forme di ripetizione, classificabili in due macrocategorie:

(a) ripetizione di enunciati di altri al fine di conferire coerenza e coesione al discorso;

(b) autoripetizione automatica utilizzata come meccanismo di controllo nella pianificazione del discorso.

Esse sono strettamente correlate alla frequenza con cui avviene il cambio dei turni: minore è la frequenza, minore sarà l'autoripetizione. Questo è dovuto alla necessità del parlante di avere il tempo necessario per pianificare e creare istantaneamente il testo durante l'interazione. È cruciale sottolineare che in vari contesti comunicativi, la ripetizione viene utilizzata come strategia stilistica per rafforzare o enfatizzare il significato del messaggio trasmesso.

- Deissi: l'integrazione nel discorso di elementi contestuali produce informazioni sufficienti per comprendere gli enunciati e consente un uso frequente di elementi deittici, come pronomi personali, dimostrativi e possessivi e indicazioni temporali ecc. La frequente presenza di deissi è causata dalla necessità di rendere lo scambio comunicativo efficiente e rapido, sfruttando al minimo gli elementi verbali. Alcune forme di ellissi possono essere interpretate come fenomeni deittici, secondo quanto indicato da Berretta nel 1994. In questo caso, si tratta di enunciati in cui l'ellissi non è strettamente legata a ciò che è stato precedentemente esposto, ma piuttosto a ciò che è già noto o presente ai partecipanti alla comunicazione.
- Prosodia. Tutti i parametri prosodici hanno un ruolo linguistico fondamentale sia a livello fonico segmentale sia a livelli quelli superiori (sintattico, semantico e testuale). Ogni parametro prosodico svolge una funzione linguistica. Per la costruzione del significato sono rilevanti il livello tonale degli enunciati, il fraseggio prosodico e i fenomeni dipendenti dal tempo del discorso. Le variazioni di tempo sono, ad esempio, legate sia alle emozioni e agli atteggiamenti dei

parlanti, sia alla distribuzione delle informazioni all'interno del testo. In effetti, si applica una velocità differente a diverse parti del testo a seconda dell'obiettivo di metterle in primo o secondo piano, o di aumentare o diminuire l'attenzione dell'ascoltatore (Savy & Cutugno, 1999). La prosodia può cambiare radicalmente la sintassi di una frase, non solo il suo valore pragmatico.

- Uso di lessico e strutture polisemiche: i testi orali presentano la più alta percentuale di parole polisemiche. La necessità di progettare e parlare contemporaneamente non consente di impiegare molto tempo per scegliere le parole e questo favorisce la selezione di strutture e parole multifunzionali e polisemiche. Si preferiscono, quindi, i sinonimi di maggiore ampiezza semantica come dimostrato dagli studi di De Mauro (et al. 1993) e Berretta (1994).
- Uso di connettivi: i testi orali prediligono l'uso di connettivi e i più frequenti sono i pronomi relativi e le congiunzioni (Voghera 2004).
- Sintassi additiva: la necessità di ricordare porzioni di testo senza il supporto di una memoria esterna determina una sintassi additiva. La sequenza delle frasi tende a riprodurre la sequenza degli eventi. Il parlato ha una sintassi di tipo concatenativo, con minor numero di subordinate e con un tipo di subordinazione che tende a conservare la progressione dell'informazione e successione delle frasi (Voghera, 2013).

2.2. Caratteristiche della lingua scritta

L'acquisizione della scrittura e della lettura è diversa da quella del parlare e comprendere il parlato. Gli usi scritti sono per definizione non nativi e non appartenenti alla varietà materna, intesa come la forma di linguaggio acquisita fin dalla nascita e che si sviluppa come il principale mezzo di pensiero e comunicazione (Voghera, 2017). La nozione di lingua materna, anche solo intuitivamente, è associata al tipo di rapporto intersoggettivamente condiviso tra una lingua e coloro che la parlano. Anche quando si impari a scrivere nella stessa lingua, non si tratta mai della stessa varietà dal punto di vista diafasico (Lepschy, 2002). La competenza nella lingua scritta è considerata, quindi, parte della socializzazione secondaria, in cui scuola e altre istituzioni svolgono un ruolo molto importante. Tuttavia, l'accesso all'istruzione e all'alfabetizzazione non è uniforme in tutto

il mondo. Di conseguenza, mentre la capacità di utilizzare la lingua parlata è un attributo comune a tutti gli individui, la competenza nella lingua scritta è limitata a una minoranza. Il fatto che la lingua scritta venga generalmente insegnata, ha portato alla convinzione che essa rappresenti il “grammaticale”, ovvero la lingua corretta, mentre molte varianti della lingua parlata sono considerate errate, brutte o maleducate (Linell, 1982). La grammatica è, quindi, spesso collegata all'istruzione formale, rappresentando qualcosa che deve essere appreso. Comunemente, è vista come una costruzione logica e astratta che si sovrappone alla lingua quotidiana, non appartenendole e ancor meno appartenendo ai parlanti (Voghera, 2017). La lingua orale, invece, è quella usata primariamente e maggioritariamente per tutta la vita, per la comunicazione verbale informale, familiare soprattutto a scopo sociale.

Storicamente, la scrittura ha costituito, all'interno delle diverse comunità, forme linguistiche riservate esclusivamente a *élites* sociali e culturali, guadagnandosi così un prestigio intrinseco. È, inoltre, per lo più associata ad usi formali, istituzionali soprattutto a scopo informativo. Questo uso dominante di un determinato canale in certe situazioni porta a un'immotivata corrispondenza tra proprietà del canale e registri, tra variazioni diamesiche e diafasiche. Quindi, l'uso di un determinato canale viene identificato con l'uso di uno specifico registro: per questo il parlato viene collegato a usi informali o addirittura trascurati e lo scritto a usi formali e accurati (Voghera 2013). I testi scritti, in particolare quelli di natura letteraria, diventano la base delle le norme linguistiche e, di conseguenza, della lingua standard. Questo avviene indipendentemente dal fatto che tali testi riflettano la lingua nativa della comunità o quella condivisa dalla maggioranza dei parlanti, come indicato da Lepschy nel 2002. La maggior parte delle grammatiche di riferimento spesso trascura l'importanza della modalità di codificazione e trasmissione nel plasmare le produzioni linguistiche. Inoltre, esse presentano la lingua come un oggetto "amodale", il quale, in realtà, coincide principalmente con le sue manifestazioni scritte (Voghera, 2017). Di conseguenza, lo scritto è il punto di riferimento essenziale per la formazione dell'identità linguistica, e in alcuni casi, anche per l'identità etnica e nazionale e i dizionari e le grammatiche svolgono un ruolo cruciale in questo processo (Voghera, 2013). Analogamente, Goody (1977) sostiene che non sia sufficiente limitarsi a una sola distinzione, *langue* vs *parole* o competenza vs performance. Egli differenzia tre concetti principali: il linguaggio (*langue*), la parola (*parole*) e la scrittura (*écriture*) e ritiene che lo studio della lingua scritta sia cruciale per comprendere la storia delle società umane, il

loro sviluppo e la loro struttura, oltre ad essere uno strumento essenziale per la conservazione e la trasmissione delle conoscenze. Inoltre, Goody (2020) sostiene che la maggior parte delle grammatiche nasce da modelli principalmente, se non esclusivamente, basati sulla lingua utilizzata dalle istituzioni, dalle classi dirigenti e, in particolare, dai letterati.

La parola, quindi, precede la scrittura filogeneticamente e ontogeneticamente, e il valore scritto è spesso secondario rispetto alla struttura della lingua parlata da un punto di vista strutturale-linguistico. Ma la scrittura è iscritta e conservata su documenti permanenti (o relativamente permanenti) e il fatto che persista nel tempo e sia statica permette di studiare meglio i testi scritti (Linell, 2004). La scrittura conferisce alle lingue una caratteristica fondamentale per la riflessione metalinguistica: le rende visibili e permanenti. Trasferendo il flusso della produzione linguistica sulla pagina, la scrittura consente di percepire la lingua al di fuori di sé stessi e fissa i suoni prodotti. Questo processo permette che l'enunciato o la parola detta, originariamente legati a un singolo e specifico evento comunicativo, diventino osservabili in modo indipendente. La scrittura crea quindi una distanza tra i parlanti e la lingua, facilitando un processo di oggettivazione delle lingue, che possono così essere osservate, studiate e analizzate (Voghera, 2017). La scrittura, dunque, aiuta a organizzare, catalogare e studiare in modo più metodico, descrivendo e teorizzando il mondo attraverso il tempo e lo spazio; consente di fare storia attraverso narrazioni critiche; agevola il progresso tecnologico; la costruzione di mondi culturali ed edifici fisici e la formulazione di prospettive individuali. I testi scritti sono, quindi, al centro delle conquiste culturali dell'umanità (Linell, 2004). L'introduzione della scrittura, secondo Farrel (1985), ha favorito una forma di pensiero più analitica e una profonda trasformazione nella struttura delle società umane. Egli suggerisce che la transizione dalla cultura orale a quella scritta abbia contribuito alla rinascita culturale e intellettuale ed evidenzia come la scrittura abbia portato a un cambiamento nella coscienza umana, influenzando la struttura del pensiero e il modo in cui le persone concepiscono il mondo e se stesse. Goody (2020) definisce la scrittura come una nuova tecnologia dell'intelletto: *“So when I use the phrase “technology of the intellect” about writing, I am thinking mainly not about the primary level of physical instrumentation but about the way that writing affects cognitive or intellectual operations, which I take in a wide sense as relating to the understanding of the world in which we live, especially the general methods we use for this”*.

Secondo Linell (2004), la comunicazione attraverso i due mezzi, parlato e scritto, è influenzata da condizioni completamente differenti. Nell'interazione dinamica del discorso parlato, caratterizzato da risposte immediate e rapide, le norme grammaticali appaiono essere di importanza secondaria. Nel discorso scritto, invece, le regole convenzionali ed esplicite hanno un impatto significativo. Secondo l'analisi di Miriam Voghera (2013), le differenze tra la lingua parlata e quella scritta risiedono nel loro diverso grado di densità e compattezza delle informazioni. Farrel (1985) sottolinea come la scrittura, introducendo una dimensione spaziale e consentendo una registrazione permanente del discorso, influenzi la percezione del tempo e dello spazio.

Per quanto riguarda la temporalità, a differenza del parlato, i testi scritti, non vengono percepiti e interpretati negli stessi tempi in cui vengono prodotti. Lo scritto è rigido e sequenziale, non offre possibilità di retroazione, non si può intervenire in corso d'opera e non si possono immaginare le possibili reazioni del destinatario. Inoltre, è regolato e programmato e solitamente non presenta margini di ambiguità (Antonelli, 2019). Un altro aspetto peculiare della lingua scritta è la sua fruibilità libera per il destinatario, che non è vincolato a una lettura lineare. Spesso i lettori non leggono per intero un testo, ma solo specifiche parti o quelle che offrono un'idea generale del contenuto complessivo.

A differenza del testo parlato, il testo scritto rivolto a una molteplicità di destinatari indifferenziati e non può basarsi sui due strategie tipiche dell'oralità descritte in precedenza: la presupposizione e la deissi. La testualità di un tipico testo scritto è molto diversa da quella propria dell'oralità e le caratteristiche intrinseche dello scritto sono:

- Lessico ricco e vario: il lessico usato nello scritto è caratterizzato da una ricchezza e varietà maggiori rispetto al parlato, si osserva una maggiore presenza di nomi e i sostantivi e le strutture di subordinazione nominale tendono ad essere proporzionalmente più frequenti (Voghera, 2005, 2013). Anche secondo Antonelli (2019) la lingua scritta si avvale dell'uso di connettivi più rigorosi e offre un lessico più ricco, sensibile alle connotazioni evocative o ironiche, rispetto a quella parlata.
- Concisione: chi scrive tende ad essere meno ridondante perché trasferisce un maggiore carico di lavoro sul lettore del testo che ha tempo per ricostruire la rete di relazioni semantico-sintattiche.

- Sintassi compatta: la lingua scritta si distingue per una sintassi più densa, in cui le valenze dei verbi vengono riempite mediante l'uso di sintagmi complessi, i quali sono connessi tra loro da preposizioni. Questo contribuisce a rendere la struttura del linguaggio più elaborata e articolata.
- Grafematica: che comprende segni e regole grafiche come interpunzione, accenti, apostrofi, differenziazione tra maiuscole e minuscole, l'uso di stili di carattere diversi come tondo e corsivo, eccetera, costituisce un aspetto peculiare e unicamente pertinente al testo scritto (Antonelli, 2019).
- Coesione: gli strumenti della coesione sono molto più articolati e vari: nel parlato il tema viene abitualmente ripetuto e i coesivi sono ridotti. Il testo scritto ricorre abitualmente a meccanismi come la riformulazione, la sostituzione lessicale mediante sinonimi, iperonimi o nomi generali, i rinvii anaforici.
- Uso dei connettivi: l'impiego dei connettivi nello scritto descrittivo e argomentativo è caratterizzato da un rigore significativamente maggiore rispetto al parlato. Nel contesto scritto, si richiede un uso più attento e formale dei connettivi, mentre nel parlato, espressioni come "insomma" o "dunque" possono essere utilizzate in maniera più informale e inerziale. È importante sottolineare che nella lingua scritta è necessario esplicitare in modo chiaro e inequivocabile la relazione di causa ed effetto che contribuisce a garantire una maggiore coesione e coerenza nel testo.
- Struttura periodale complessa: la lingua scritta è caratterizzata dall'uso diffuso di periodi complessi e di molte subordinate. Questa caratteristica riflette una sintassi più elaborata (Antonelli, 2019).

2.3. Bias della lingua scritta

Nessun ricercatore, e nessun singolo studio, può concentrarsi su tutto, quindi, sono inevitabili dei "bias". Le seguenti definizioni di "bias" possono essere trovate nell'Oxford Advanced Learner's Dictionary (2010):

bias 1: *“a strong feeling in favour of or against one group of people, or one side in an argument, often not based on fair judgement”*, 2: *“an interest in one thing more than others: a special ability (.)”*; bias v 1: *w sb/sth: “to unfairly influence sb’s opinions or*

decisions (syn: prejudice)”; 2: w sth: “to have an effect on the results of research or an experiment so that they do not show the real situation (.)”

Un “bias” in una teoria, spiegazione o descrizione è uno squilibrio di interesse scientifico, tipicamente con un focus su alcune forme, funzioni o fenomeni, e una quasi completa negligenza di un altro dominio o di tutti gli altri domini (Linell, P. 2004).

Linell (1982, 2004, 2012) ha identificato un *Written Language Bias* nella linguistica, sostenendo che nonostante il riconoscimento della lingua parlata come primaria sotto molti aspetti, cioè come rappresentazione naturale e vera della lingua, i modelli e metodi per la descrizione della lingua parlata e della lingua in generale sono stati ripresi dalla tradizione di descrivere e spiegare la lingua scritta. Da sempre ci si è occupati della lingua scritta in misura maggiore rispetto a quella parlata, per diversi motivi:

- a) La progressione tecnologica del mezzo della lingua scritta ha significativamente influenzato la nostra percezione e comprensione della natura della lingua in generale.
- b) Le ragioni e i fini che hanno guidato l'attività dei linguisti nel passato. Questi obiettivi non mantengono più la stessa preminenza oggi, ma le direzioni tradizionali che la linguistica ha seguito, caratterizzano ancora le molte teorie e pratiche contemporanee;
- c) L'elevato status della lingua scritta in quasi tutte le società e quindi anche tra i linguisti appartenenti a quelle società;
- d) Il ruolo della lingua scritta come metalinguaggio della descrizione linguistica. La lingua scritta non era solo oggetto di analisi scientifica, ma era anche (ed è tuttora) il mezzo in cui venivano enunciati i prodotti dell'analisi, le diverse teorie e posizioni metalinguistiche.

Il *bias* della lingua scritta riguarda, quindi, non solo teorie e metodi nei risultati dei linguisti, nella teoria linguistica, ma è anche presente nel senso comune del linguaggio tra le persone e in generale nella cultura occidentale. Linell pone l'attenzione sui modi in cui la linguistica tende a trattare le relazioni tra la struttura e l'uso del linguaggio nel parlato e nello scritto. Secondo lui, i linguisti hanno ignorato le differenze tra i sistemi linguistici alla base del parlato e della scrittura. C'è un'evidente contraddizione tra ciò che i linguisti ritengono riguardo ai rapporti tra parola e scrittura, e ciò che effettivamente

fanno nella pratica. La teoria del *Written Language Bias* (Linell, 2004) sostiene che la linguistica sia finita in un paradosso; poiché proclama il primato della parola e del linguaggio parlato, eppure rimane in gran parte fedele ai metodi e ai modelli tradizionali sviluppati per lo studio della lingua scritta. Da un lato, i linguisti sostengono che la lingua parlata sia primaria e, quindi, l'oggetto di studio più interessante. D'altro canto, la loro pratica concreta rivela tradizioni e valori piuttosto diversi. In larga misura, i linguisti studiano ancora le frasi scritte (spesso fuori contesto) e i testi scritti, e lo studio della lingua parlata viene fatto relazionandola alla lingua scritta. Inoltre, tendono a ignorare le differenze tra le due, presumendo che sia presente un unico sistema, quello che Saussure chiamava “*la langue*”, che sta alla base sia del parlato che dello scritto. Chomsky dà per scontato che esista una competenza linguistica monolitica, non un aggregato di capacità distinte, una per la lingua usata per parlare e ascoltare e una per la lingua usata per scrivere e leggere; non sembra aver mai visto alcun motivo per concentrarsi sulle importanti differenze tra parlato e scrittura.

Tuttavia, se si guarda ai fondatori della linguistica strutturalista, è facile trovare affermazioni come le seguenti:

« *Langage et écriture sont deux systèmes de signes distinct l'unique raison d'être du second est de représenter le premier* » (Saussure 1964)

“*Writing is not language, but merely a way of recording language by means of visible marks*” (Bloomfield 1933).

Quindi, Bloomfield, la scrittura è semplicemente una “rappresentazione secondaria” del discorso. Questi punti di vista sono condivisi dalla maggior parte dei linguisti. Per esempio, Hockett, nel suo famoso libro (*A Course in Modern Linguistics*, 1958), distingue tra “linguaggio” e “scrittura”, sostenendo che “*language excludes writing*”; ed è evidente che per “*language*” si riferisca alla lingua parlata. Anche Lyons sostiene che “ *it is one of the cardinal principles of modern linguistics that spoken language is more basic than written language* ” (Lyons 1981: 11). La scrittura solitamente è considerata come una rappresentazione meno problematica e teoricamente meno controversa della lingua orale. Questa opinione è stata sostenuta a lungo, infatti, anche secondo Rousseau « *Les langues sont faites pour être parlées, l'écriture ne sert que de supplément à la parole ... L'écriture n'est que la représentation de la parole, il est bizarre*

qu' on donne plus du soin à déterminer l' image que l' objet. » (Rousseau (a cura di Derrida) 1967)

Inoltre, un sistema di scrittura è definito come la “rappresentazione visiva della lingua parlata su carta o altro supporto” (Bird et al., 2022). Secondo Thormoset (2006) una lingua per essere definita scritta dovrebbe avere almeno un sistema di scrittura a livello locale. Tuttavia, afferma che dovrebbero essere presenti più materiali per distinguere una lingua scritta da una lingua non scritta, come dizionari, testi, letteratura e materiali per sostenere un programma di alfabetizzazione continuo.

È fondamentale comprendere che quando ci si riferisce alle "lingue scritte", ci si riferisce a un sistema linguistico che non è limitato esclusivamente alla forma scritta, ma che ha una presenza orale altrettanto significativa. La nozione di "lingua scritta" si basa sull'idea che le lingue sono sistemi complessi e dinamici che possono manifestarsi sia attraverso la modalità scritta che attraverso quella orale. Il concetto di "lingue scritte", quindi, riconosce l'interconnessione intrinseca tra la modalità orale e quella scritta di una lingua. Una visione completa di una lingua considera entrambe le modalità come parte integrante del sistema linguistico, riflettendo la complessità e l'adattabilità delle lingue umane nel loro insieme (Voghera, 2017). Inoltre, poiché la lingua parlata e quella scritta si influenzano reciprocamente in modo costante, è comune che la lingua scritta prenda elementi dalla forma parlata e viceversa (Antonelli, 2019). Nonostante questo, le caratteristiche che distinguono la comunicazione parlata dalla scritta sono poco comprese. Quindi, secondo Linell, oggi è necessaria un'analisi più accurata della lingua parlata e della lingua scritta. Negli ultimi decenni, lo studio della variazione diamesica è diventato più popolare nel campo della linguistica, anche grazie agli studi di Norbert Dittmar (Dittmar, Bahlo (a cura di) 2010), Peter Auer (2009), Miriam Voghera (2005, 2010, 2013, 2017). La variazione diamesica si concentra principalmente sulle variazioni linguistiche che si verificano a causa di un cambiamento nel mezzo di trasmissione, ovvero nel canale di comunicazione. Tuttavia, anche secondo gli studi più recenti di Miriam Voghera (2013) molto raramente il cambio di canale è un mero passaggio di sistema di trasmissione, comporta invece variazioni di contesto, destinatari e temi e argomenti, che sono collegati a registri linguistici differenti. L'analisi dettagliata di queste variazioni rappresenta una necessità cruciale nell'ambito dell'analisi linguistica odierna.

3. Materiali e metodi

3.1. Risorse utilizzate

3.1.1. *World Atlas of Language Structures (WALS)*

La principale risorsa disponibile online per osservare la diversità linguistica è il *World Atlas of Language Structures* (WALS.info, Dryer & Haspelmath 2013), che è un ampio database che racchiude le proprietà strutturali (fonologiche, grammaticali, lessicali) delle lingue raccolte, ovvero 2.560, basandosi su materiali descrittivi (come le grammatiche di riferimento) creato da un team di 55 autori. La prima versione di WALS è stata pubblicata come libro con CD-ROM nel 2005 dalla *Oxford University Press*. La prima versione online è stata pubblicata nell'aprile 2008. L'edizione 2013 di WALS corregge una serie di errori di codifica. A partire dall'edizione 2013 di WALS, vengono rilasciate e pubblicate periodicamente serie di correzioni. WALS Online è una pubblicazione del *Max Planck Institute for Evolutionary Anthropology*. Si tratta di una pubblicazione separata, a cura di Dryer, Matthew S. e Haspelmath, Martin (Lipsia: Max Planck Institute for Evolutionary Anthropology, 2013). Il programmatore principale è Robert Forkel.

Il WALS è composto da 142 mappe con testi esplicativi sulle diverse caratteristiche (come la dimensione dell'inventario delle vocali, l'ordine dei sostantivi e del genitivo, le costruzioni passive e la polisemia di "mano"/"braccio"). Ciascuna mappa mostra tra 120 (35) e 1110 lingue, ciascuna lingua è rappresentata da un punto e diversi colori di punto mostrano diversi valori delle caratteristiche. Sulle mappe vengono visualizzate complessivamente 2.650 lingue e più di 58.000 punti forniscono informazioni sulle caratteristiche di determinate lingue. Questa organizzazione di dati in forma di atlante consente agli utenti di visualizzare e confrontare facilmente le caratteristiche linguistiche di diverse lingue. L'uso di mappe e grafici facilita la comprensione delle distribuzioni geografiche e delle correlazioni tra le caratteristiche linguistiche. Il WALS utilizza due campioni distinti per analizzare le caratteristiche linguistiche delle lingue del mondo: un campione principale di 100 lingue e un campione esteso di 200 lingue. Questi campioni sono stati selezionati seguendo criteri specifici per garantire rappresentatività e varietà geografica. Entrambi i campioni cercano di

rappresentare in modo equo la diversità geografica delle lingue del mondo. Quindi, le lingue sono selezionate in modo da coprire diverse regioni globali, inclusi continenti e aree linguistiche significative. Sono state incluse lingue provenienti da diverse famiglie linguistiche per riflettere la varietà di alberi genealogici linguistici. Ciò consente di esaminare non solo le differenze tra le lingue, ma anche le somiglianze e le divergenze all'interno delle famiglie linguistiche. Per il campione esteso di 200 lingue è stata inclusa una maggiore varietà di lingue per consentire ulteriori approfondimenti. La scelta specifica di lingue nei campioni può variare nel tempo a seconda degli aggiornamenti del WALS e delle nuove informazioni linguistiche disponibili. Infatti, WALS è soggetto a aggiornamenti regolari per incorporare nuovi dati e riflettere gli sviluppi nella ricerca linguistica. Questo assicura che le informazioni fornite siano sempre aggiornate e rilevanti.

Il WALS è un progetto a disposizione del pubblico e offre un accesso aperto ai dati sulla diversità strutturale delle lingue del mondo attraverso il suo sito web. È uno strumento chiave nel campo della linguistica comparativa e della documentazione linguistica. Attraverso la sua presentazione organizzata di dati linguistici e la facilità di accesso, contribuisce notevolmente alla comprensione della diversità delle strutture linguistiche nelle lingue del mondo.

3.1.2. Ethnologue: Languages of the World

Ethnologue: Languages of the World è una pubblicazione cartacea ed elettronica del SIL International (conosciuta come Summer Institute of Linguistics), un'organizzazione cristiana non profit, che studia e documenta essenzialmente le lingue meno conosciute, particolarmente per far arrivare il messaggio della Bibbia al maggior numero possibile di lettori nella loro madrelingua. Ethnologue è una delle risorse più ampie e complete per la documentazione delle lingue del mondo. Nella diciottesima edizione di Ethnologue, pubblicata nel 2015, sono state fornite statistiche su 7.472 lingue e per ciascuna di queste fornisce informazioni dettagliate, inclusi dettagli sulla struttura linguistica, l'estensione geografica, il numero dei parlanti, la classificazione linguistica, le affiliazioni linguistiche, la reperibilità della Bibbia e altro ancora, ma non da informazioni statistiche sul numero di lingue che hanno una forma scritta, viene fornito il tipo di scrittura utilizzato per i materiali scritti, se noto. Ethnologue classifica le lingue in famiglie

linguistiche e gruppi, aiutando a identificare le relazioni genetiche tra le lingue. Questa classificazione offre una panoramica della diversità linguistica e delle connessioni tra le lingue. Fornisce informazioni sulla distribuzione geografica delle lingue, consentendo agli utenti di esplorare dove sono parlate le lingue specifiche e quali comunità le utilizzano. Ethnologue include stime della popolazione parlante ogni lingua e offre informazioni demografiche sulle comunità etnolinguistiche. Ciò aiuta a comprendere la vitalità delle lingue e le sfide che possono affrontare. Documenta le variazioni linguistiche all'interno di una lingua, comprese le varietà dialettali e le differenze regionali. Questo è cruciale per comprendere la ricchezza e la complessità delle lingue.

Ethnologue è utilizzato da linguisti, antropologi, studiosi di studi culturali e altri professionisti del settore come fonte affidabile di informazioni linguistiche. È una risorsa di riferimento per chiunque sia coinvolto nella ricerca etnolinguistica. Il progetto contribuisce alla conservazione linguistica fornendo informazioni che possono essere utilizzate per preservare e documentare lingue a rischio di estinzione. È soggetto ad aggiornamenti regolari per riflettere gli sviluppi nella ricerca linguistica e le nuove informazioni disponibili sulle lingue del mondo. In sintesi, Ethnologue è una risorsa chiave per la documentazione, la classificazione e la comprensione delle lingue globali, contribuendo alla conservazione e alla promozione della diversità linguistica a livello globale e ad oggi, è l'inventario di lingue più esauriente.

La definizione di "una lingua" su Ethnologue non si limita ai soli criteri linguistici. All'interno della risorsa, viene utilizzato lo standard ISO 639-3 identifica univocamente ogni lingua ed è utilizzato anche per costruire l'url della pagina di ogni lingua sul sito. Il codice ISO 639-3 applica i seguenti criteri fondamentali per definire una lingua in relazione a varietà che potrebbero essere considerate dialetti:

- Due varietà correlate sono considerate parte della stessa lingua se i parlanti di ciascuna varietà comprendono l'altra varietà linguistica (ovvero, possono capirsi reciprocamente sulla base della conoscenza della propria varietà linguistica senza dover imparare l'altra varietà).
- Se l'intelligibilità tra due varietà linguistiche è marginale, hanno una letteratura o un'identità etnolinguistica comune con una varietà centrale compresa da entrambe le comunità di parlanti, queste dovrebbero essere considerate varietà linguistiche della stessa lingua individuale.

- Se c'è sufficiente intelligibilità tra due varietà per consentire la comunicazione, queste possono comunque essere trattate come lingue diverse quando hanno identità etnolinguistiche diversamente nominate da lungo tempo, abbinata a standardizzazione e letterature distinte.

Per la creazione dei campioni, sono stati esaminati determinati criteri all'interno di Ethnologue:

- **Language Development:** fornisce informazioni sui tassi di alfabetizzazione, sull'uso nell'istruzione, sulla documentazione e sui prodotti dello sviluppo linguistico, sugli sforzi di rivitalizzazione e sugli enti che si occupano dello sviluppo linguistico di quella data lingua. In particolare, per la creazione dei campioni ci si è concentrati su:
 - Percentuali della popolazione di parlanti che sono alfabetizzati per le lingue L1 e L2.;
 - Uso nelle scuole primarie o secondarie. La lingua può essere utilizzata come lingua di insegnamento o insegnata come materia in una o più scuole nell'area linguistica.
 - Pubblicazioni e utilizzo nei media: Viene indicata l'esistenza di materiali prodotti nella lingua, come documentazione linguistica (dizionari, grammatiche, testi), letteratura stampata e media trasmessi, quando noti. Quando esiste una vasta letteratura e produzione media, la lingua è identificata come "*Fully developed*".
- **Writing:** per ciascuna lingua, viene indicato il sistema di scrittura e gli alfabeti utilizzati per i materiali scritti, se noti. Sono riportati anche eventuali stili specifici di uno script utilizzato, gli anni in cui uno script ha iniziato o cessato di essere utilizzato, e altri commenti riguardanti la scrittura e l'ortografia. In generale, quando non viene identificato alcun sistema di scrittura, si può presumere che non esista un sistema di scrittura ampiamente accettato e utilizzato.
- **Language Use**
 - *Language status:* viene riassunto lo stato di ogni lingua in ciascun paese in cui è utilizzata, fornendo due tipi di informazioni. La prima è una stima dello sviluppo complessivo o della minaccia della lingua utilizzando la scala EGIDS (Lewis e Simons 2010). Il secondo è una categorizzazione del riconoscimento ufficiale attribuito a una lingua all'interno del paese.

La scala EGIDS è composta da 13 livelli, con ogni numero più alto sulla scala che rappresenta un maggiore livello di interruzione nella trasmissione intergenerazionale della lingua. In linea generale, quando la lingua viene trasmessa ai bambini come loro prima lingua, o quando viene utilizzata frequentemente e ampiamente all'interno della comunità, si utilizza più spesso il termine "Vigorous" (Vivace).

- *Domains of use*: Quando più di una lingua viene utilizzata in una comunità, i parlanti spesso stabiliscono specifici interlocutori, argomenti e luoghi.
- *User age range*: questo campo descrive la fascia d'età di coloro che utilizzano la lingua come L1. A mano a mano che avviene il cambiamento linguistico, gli adulti più anziani tendono ad essere gli ultimi parlanti della lingua tradizionale.
- ***Literacy rate***: si tratta di una stima percentuale della popolazione del paese che è alfabetizzata in una qualsiasi lingua. I dati provengono principalmente dall'UNESCO, ma possono derivare da varie altre fonti se sono disponibili stime più recenti.
- ***User population***: fornisce il numero di persone nel paese che utilizzano questa lingua, oltre al numero totale di utenti in tutto il mondo se la lingua è utilizzata in più paesi. Queste popolazioni utenti sono suddivise tra utenti di prima e seconda lingua quando i dati sono disponibili. Possono essere inclusi anche dati sulla popolazione monolingue, sulla popolazione etnica e altri commenti sulla popolazione.

3.1.3. ***Grambank***

Grambank è stato costruito in una collaborazione internazionale tra gli istituti Max Planck di Lipsia e Nimega, l'Università Nazionale Australiana, l'Università di Auckland, l'Università di Harvard, l'Università di Yale, l'Università di Turku, l'Università di Kiel, l'Università di Uppsala, SOAS, Endangered Languages Documentation Program e oltre un centinaio di studiosi da tutto il mondo. Grambank è stato progettato per essere utilizzato per indagare la distribuzione globale di caratteristiche, universali linguistici, dipendenze funzionali, preistoria della lingua e interazioni tra lingua, cognizione, cultura e ambiente. Il database Grambank copre attualmente 2.467 varietà linguistiche,

catturando un'ampia gamma di fenomeni grammaticali in 195 caratteristiche, dall'ordine delle parole al tempo verbale, ai plurali nominali e molte altre variabili linguistiche comparative ben studiate. Grambank copre 215 famiglie linguistiche diverse e 101 lingue isolate da tutti i continenti abitati. L'obiettivo è che Grambank copra in definitiva tutte le lingue per le quali esiste una grammatica o una bozza di grammatica. Grambank fa parte di Glottobank, un consorzio di ricerca che prevede il lavoro su database complementari di dati lessicali, paradigmi, numeri e modelli sonori nelle lingue del mondo. Grambank può essere utilizzato insieme ad altri database, come quelli di Glottobank e D-PLACE, per approfondire la nostra comprensione della nostra storia e delle nostre capacità comunicative.

3.1.4. Glottolog

Glottolog è un progetto linguistico online sviluppato e mantenuto *dall'Istituto Max Planck per la scienza della storia umana*. Glottolog fornisce un catalogo completo e accessibile pubblicamente che contiene informazioni sulle lingue, le famiglie linguistiche e i dialetti del mondo. Assegna un identificatore unico e stabile (il Glottocode) a tutte le famiglie, lingue e dialetti. Qualsiasi varietà su cui lavora un linguista ottiene prima o poi una propria voce. Le lingue sono organizzate attraverso una classificazione genealogica (un albero) che si basa sulle ricerche storico-comparative disponibili. Glottolog fornisce dati sulla struttura grammaticale, fonologica e lessicale di molte lingue. Gli studiosi possono accedere a risorse linguistiche come descrizioni grammaticali, dizionari e altre fonti utili. Inoltre, include una vasta raccolta di riferimenti bibliografici relativi alle lingue e alle loro caratteristiche. Glottolog serve come una risorsa fondamentale per chiunque sia interessato allo studio e alla ricerca delle lingue del mondo. Gli studiosi, i linguisti e gli studenti possono utilizzare il sito per accedere a dati affidabili e aggiornati sulla classificazione, la struttura e altre caratteristiche linguistiche di molte lingue diverse.

3.2. Procedura di costruzione dei campioni

In questa tesi, è stato preso in esame un campione di 200 lingue presente nel WALS nella sezione “*Languages*”. Utilizzando questo campione come riferimento, sono stati creati due ulteriori campioni. Il primo comprende lingue esclusivamente orali, prive di una

tradizione scritta, mentre il secondo include lingue che, oltre ad essere lingue orali, hanno anche una consolidata tradizione scritta. Nel processo di costruzione dei due campioni è stato fondamentale l'uso di diversi altri database, tra cui Ethnologue e Glottolog. Questi strumenti hanno permesso di sviluppare campioni bilanciati che rispecchiassero le caratteristiche geografiche e genealogiche presenti nel campione originale del WALS.

Per la creazione dei due campioni linguistici, si è condotta un'analisi approfondita delle risorse presenti nel database di Ethnologue per ogni lingua inclusa nel campione originale del WALS. L'obiettivo era comprendere se ciascuna lingua possedesse o meno una forma scritta. Si è proceduto a un'attenta selezione delle lingue presenti nel campione iniziale del WALS, ovvero a un esame individuale di ciascuna lingua, l'inserimento del relativo codice ISO fornito dal WALS nella piattaforma di Ethnologue e la definizione di parametri per distinguere quali lingue potessero essere classificate come principalmente orali e quali come lingue con una robusta tradizione scritta. A tale scopo, sono stati applicati specifici criteri:

- se la lingua contiene "*Unwritten*" nel campo "*Writing*" di Ethnologue, è inserita nel campione di lingue orali.
- se il campo "*Language Development*" di Ethnologue contiene le parole "*Literature*" oppure "*Newspapers*" oppure "*TV*" oppure "*Radio*", la lingua è classificata come lingua con tradizione scritta.
- se nel campo "*Language Development*" di Ethnologue viene specificato che la lingua è utilizzata come lingua di insegnamento, questa è inserita nel campione di lingue scritte.
- se il campo "*Writing*" contiene almeno una delle parole "*NT*", "*OT*", "*Bible*", "*Dictionary*", "*Grammar*" ma non contiene nessuna delle parole "*Literature*", "*Newspaper*", "*Radio*", "*TV*", allora la lingua viene inserita nel campione di lingue orali.
- se la lingua è identificata come "*Fully developed*" nel campo "*Language Development*", è stata inserita nel campione di lingue scritte.

Per la classificazione delle lingue prive di informazioni adeguate all'interno delle sezioni elencate, è stata effettuata un'analisi delle "*Language resources*" che forniscono un collegamento alla pagina del catalogo dell'Open Language Archives Community

(OLAC) che elenca risorse nella lingua e relative ad essa, al fine di valutare la tipologia dei testi scritti presenti. In questa sezione di Ethnologue le risorse scritte sono suddivise in “*primary texts*”, “*language description*” e “*other resources*”.

- In caso di presenza di “*primary texts*”, si è andata a verificare la quantità di questi testi e soprattutto se questi fossero o meno testi in lingua originale o traduzioni della bibbia e testi sacri. Quindi le lingue i cui primary text erano esclusivamente traduzioni della bibbia e testi sacri sono state inserite nel campione di lingue orali. Le altre, nel campione di lingue scritte.
- Le “*language descriptions*” non sono state considerate parte della tradizione scritta poiché possono consistere in grammatiche o raccolte lessicali o in dettagliate analisi e descrizioni della lingua fatte da studiosi. Pertanto, le lingue che presentavano solamente “*language descriptions*” sono state inserite nel campione di lingue orali.

Nei casi in cui i dati nelle “*external resources*” erano insufficienti, ulteriori informazioni sono state ricercate attraverso il motore di ricerca Google, ma è stato cruciale anche l'utilizzo di risorse provenienti da siti web e biblioteche online per acquisire informazioni supplementari. Ad esempio, è stato fatto ricorso al catalogo della *National Library of Australia*, una risorsa che raccoglie una vasta gamma di materiali quali libri, testi, immagini, mappe, interviste di storia orale, manoscritti e altre fonti relative alle lingue e ai dialetti australiani.

Nel corso dell'analisi, per il campione di lingue orali, se una lingua appartenente al WALIS risultava avere una tradizione scritta consolidata, si è cercato, quando possibile, di sostituirla con una lingua della stessa famiglia linguistica che, al contrario, non presentasse una tradizione scritta consolidata. Viceversa, per il campione di lingue scritte, se una lingua non aveva una tradizione scritta consolidata, si è cercato di sostituirla con una lingua appartenente alla stessa famiglia linguistica ma con una tradizione scritta consolidata. Questa pratica di scambio ha contribuito a creare campioni bilanciati che riflettono le caratteristiche specifiche desiderate per l'analisi.

Infine, i due campioni sono stati raccolti in una tabella che comprende diverse categorie di dati: il nome della lingua, il genere, la famiglia linguistica di appartenenza, la localizzazione geografica in cui è parlata, nonché i contenuti, inclusi sulla base delle fonti quali Ethnologue. Tra i contenuti considerati vi sono elementi quali la presenza di

letteratura, giornali, trasmissioni radiofoniche, televisive, testi sacri come la Bibbia, dizionari, grammatiche, e ulteriori informazioni rilevanti, come ad esempio la prossimità linguistica con altre lingue.

Al fine di verificare il bilanciamento geografico e genealogico, è stata condotta un'analisi delle famiglie linguistiche presenti all'interno dei tre campioni, incluso quello originario del WALS. Questo approccio ha permesso di calcolare con precisione il numero di famiglie linguistiche rappresentate in ciascun campione, fornendo un quadro dettagliato e accurato delle relazioni genealogiche e geografiche presenti all'interno delle varie tipologie di lingue considerate.

Per valutare la veridicità dei campioni creati e confermare o meno la presenza di testi scritti è stato poi analizzato il *Database* descritto nel capitolo successivo.

3.2.1. Focus sulle criticità- Lingua Maung

Nel campione iniziale di 200 lingue del WALS è inclusa la lingua Maung (ISO mph). Tuttavia, in specifiche circostanze, come nell'esempio citato, le informazioni fornite da Ethnologue non risultano sufficienti per effettuare un giudizio accurato sulla presenza o l'assenza di una tradizione scritta, come mette in luce l'immagine sottostante. Ethnologue segnala la presenza di testi, ma non menziona una vera e propria letteratura associata alla lingua Maung. Inoltre, la carenza di dettagli sullo sviluppo linguistico o sulla alfabetizzazione è particolarmente evidente, accentuata ulteriormente dal fatto che la lingua è prossima all'estinzione (8b), gli unici utenti rimasti della lingua sono membri della generazione dei nonni o più anziani, i quali hanno limitate occasioni di utilizzare la lingua.

ISO 639	amg
Alternate Names	A'moordiyu, Amarag, Amarak, Amardak, Amurag, Amurdag, Amurrag, Amurrak, Amurtak, Amuruk, Mamurug, Monobar, Namurug, Nga:mr:rak, Nga:mu:rak, Ngamurag, Umoreo, Umoriu, Umorrdak, Wardadjbak, Woraibug, Wureidbug
Population	1 (2015). No reports of the death of the last known speaker, Charlie Mungulda, in 2021.
Location	Northern Territory: Goulburn island, Oenpelli.
Language Maps	Northern Australia
Language Status	8b (Nearly extinct).
Classification	Australian, Yiwaidjan, Amaragic
Language Use	Shifted to English [eng].
Language Development	Dictionary. Texts.
Digital Support	Still.
Language Resources	OLAC resources in and about Amurdak

Figura 1- Lingua Maung

All'interno delle risorse di OLAC, si trovano numerosi "*primary texts*", ma è importante notare che questa abbondanza potrebbe risultare fuorviante. In realtà, un'analisi più approfondita di queste risorse, unita all'esame del catalogo della *National Library of Australia*, ha consentito di ottenere una comprensione più dettagliata della natura dei testi presenti nella lingua in questione, come chiaramente illustrato nella figura 2 e 3. Risulta particolarmente evidente che i testi di traduzione della Bibbia o le trascrizioni di dialoghi occupino una posizione predominante. La notevole abbondanza di tali testi insinua chiaramente che questa lingua viene prevalentemente impiegata in forma orale.

Primary texts

1. [ONLINE Injalak artists and elders Jimmy Kalariya and Don Namundja discuss media from 1948 American-Australian Scientific Expedition to Arnhem Land](#). Reuben Brown (compiler); Reuben Brown (recorder); Reuben Brown (researcher); Reuben Brown (speaker); Isaiah Nagurrurrba (speaker); Jimmy Kalariya (speaker); Don Namundja (speaker). 2011. Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC). [oai:paradisec.org.au/RB2-20110802](http://oai.paradisec.org.au/RB2-20110802)
2. [ONLINE Jimmy Kalariya and family discuss media from 1948 American-Australian Scientific Expedition to Arnhem Land](#). Reuben Brown (compiler); Reuben Brown (recorder); Reuben Brown (researcher); Isaiah Nagurrurrba (speaker); Jimmy Kalariya (speaker). 2011. Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC). [oai:paradisec.org.au/RB2-20110806_01](http://oai.paradisec.org.au/RB2-20110806_01)
3. [ONLINE Jacob Nayinggul and Connie Nayinggul discuss media from 1948 American-Australian Scientific Expedition to Arnhem Land](#). Reuben Brown (compiler); Reuben Brown (recorder); Reuben Brown (researcher); Jacob Nayinggul (speaker); Connie Nayinggul (speaker). 2011. Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC). [oai:paradisec.org.au/RB2-20110806_02](http://oai.paradisec.org.au/RB2-20110806_02)
4. [ONLINE Daisy Guymala and family discuss media from 1948 American-Australian Scientific Expedition to Arnhem Land](#). Reuben Brown (compiler); Amanda Harris (speaker); Reuben Brown (speaker); Isaiah Nagurrurrba (speaker); Daisy Guymala (speaker); Virginia Nadjamerrek (speaker); Allan Nadjamerrek (speaker). 2011. Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC). [oai:paradisec.org.au/RB2-20110817](http://oai.paradisec.org.au/RB2-20110817)
5. [ONLINE Jimmy Kalariya and family discuss media from 1948 American-Australian Scientific Expedition to Arnhem Land](#). Reuben Brown (compiler); Amanda Harris (interviewer); Allan Marett (interviewer); Reuben Brown (interviewer); Reuben Brown (recorder); Jimmy Kalariya (speaker); Daisy Guymala (speaker). 2011. Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC). [oai:paradisec.org.au/RB2-20110818](http://oai.paradisec.org.au/RB2-20110818)

Figura 2- OLAC Resources

Description:	Canberra : British and Foreign Bible Society in Australia, 1969 [8] p. : ill. ; 14 cm.
Notes:	<ul style="list-style-type: none"> • This passage is St. Luke 2:1:20 in Maung"--Back cover. • Translated in Maung by Miss H. Hinch for a group of a few hundred Aborigines living in Arnhem Land, Northern Territory, Australia.--Accompanying note.
Subject:	Maung language -- Texts
Other authors/contributors:	Hinch, H, translator British and Foreign Bible Society in Australia

Figura 3- National Library of Australia

3.3 Distribuzione dei fenomeni nei campioni

3.3.1 Creazione database

- Step 1 - Analisi Ethnologue: attraverso uno script python è stato possibile ottenere una tabella in cui, per ogni lingua, si hanno i seguenti parametri:
 - ID, ISO 639, Alternate Names, Autonym, Population, Location, Language Maps, Language Status, Classification, Dialects, Language Use, Language Development, Digital Support, Writing, Other Comments, Language Resources, Typology, Member Languages
- Step 2 - "Divisione in *primarily oral*, *written*, *don't know*": sono stati stabiliti dei criteri per determinare, a partire dalle informazioni di Ethnologue, quali lingue possano ritenersi come lingue primariamente orali, altre come lingue con una forte tradizione scritta e altre come casi intermedi da eventualmente classificare a mano.

Per farlo, i criteri che sono stati applicati sono:

- se la lingua contiene "*Sign*" nel nome, quindi una lingua segnata, è classificata come "*primarily oral*"
- se contiene "*Unwritten*" nel campo "*Writing*" di Ethnologue, è classificata come "*primarily oral*"
- se il campo "*Writing*" di ethnologue è vuoto, è classificata come "*primarily oral*"

- se il campo "*Language Development*" di Ethnologue contiene le parole chiave "*Literature*" oppure "*Newspapers*" oppure "TV" oppure "Radio", è classificata come "*written*"
 - Se il campo "*Language Development*" di Ethnologue contiene la stringa "*Taught in*", è classificata come "*written*"
 - se il campo "*Language Use*" di Ethnologue contiene le parole chiave "*Shifting*" o "*Shifted*" la lingua viene classificata come "*Primarily oral*"
 - se il campo "*Writing*" contiene almeno una delle parole chiave "NT", "OT", "*Bible*", "*Dictionary*", "*Grammar*" ma non contiene nessuna delle parole "*Literature*", "*Newspaper*", "*Radio*", "*TV*", allora la lingua viene classificata come "*primarily oral*"
 - se nessuno di questi casi si applica, la lingua è classificata come "*don't know*"
- Step 3 - *Mapping glottocode – isocode*: a partire da glottolog sono stati mappati gli *iso-codes* a *glottocode*. In questo modo si conoscono il *glottocode* di molte delle lingue presenti in Ethnologue in quanto Ethnologue usa gli *iso-codes* per identificare le lingue
 - Step 4 - *Mapping WALS e GramBank*: con lo stesso principio sono stati mappati i dati da WALS e GramBank. Per ognuna di queste risorse infatti vengono forniti 4 file:
 - `languages_[wals|grambank].csv`
 - `parameters_[wals|grambank].csv`
 - `codes_[wals|grambank].csv`
 - `values_[wals|grambank].csv`
 - Nei file "*languages*" ci sono varie informazioni per ogni lingua, e più nello specifico l'ID specifico della risorsa, e per Grambank il *glottocode*, mentre per WALS sia il *glottocode* che uno o più *iso-codes*

- Nel file “*parameters*” sono descritti la lista di parametri esaminati con un ID unico e il nome, che è quello che poi è stato riportato nelle colonne nella tabella finale (es. per la colonna “GP-*Are there definite or specific articles?*”, “*Are there definite or specific articles*” è il nome di un parametro)
- Nel file “*values*” è mappato, per ogni ID parametro, i possibili valori che questo può assumere
- Il file “*codes*”, infine, riporta per ogni ID lingua per ogni ID parametro il valore assegnato

Dai mapping è emerso che i codici ISO e i codici Glottolog non hanno una corrispondenza 1:1. In questi casi, è stata duplicata l'informazione in modo da associare determinati parametri che erano stati ad esempio associato a un solo codice ISO, a tutte le lingue con lo stesso ISO ma diversi codici Glottolog o viceversa.

3.3.2. *Analisi database*

Dopo aver creato il database precedentemente menzionato, sono state estratte le lingue in esso contenute, corrispondenti alle lingue presenti nei due campioni, al fine di confermare la presenza o assenza di testi scritti. Inoltre, è stata effettuata un'analisi delle lingue del campione del WALS per verificare il bilanciamento tra lingue scritte e lingue orali.

Per esaminare la distribuzione di alcune caratteristiche linguistiche in entrambe le modalità comunicative, sono stati identificati sei fenomeni all'interno di Grambank e uno all'interno di WALS, i quali sono stati successivamente soggetti a un'analisi dettagliata. I fenomeni presenti in Grambank includono:

- **GB020-*Are there definite or specific articles?***

Questo fattore fa riferimento alla presenza o assenza di articoli definiti o specifici nella struttura grammaticale di una lingua. A volte, tali concetti di specificità e definizione sono riassunti nel termine 'identificabilità'. L'espressione formale è irrilevante; gli articoli possono essere liberi, vincolati o contrassegnati da segni sovrasegmentali come il tono. Gli articoli sono diversi dai dimostrativi in quanto i dimostrativi si verificano in un paradigma di marcatori che hanno una chiara funzione deittica spaziale. Gli articoli definiti, come "il" in italiano, indicano un sostantivo in modo specifico, identificando un oggetto o un concetto noto o già menzionato nel contesto della comunicazione.

Grambank da un valore “1” se c'è un morfema che può marcare la definizione o la specificità senza trasmettere anche un significato deittico spaziale, valore “0” se la fonte non menziona un articolo definito e non si trovano esempi o testi in una grammatica esaustiva e valore “?” se la grammatica non fornisce un'analisi sufficiente per determinare se esiste un articolo definito o meno. All'interno del database il valore “1” di Grambank è indicato come “presente” e il valore “0” come assente.

- **GB021-Do indefinite nominals commonly have indefinite articles?**

Questo fattore si riferisce alla frequenza con cui i nomi indefiniti sono accompagnati da articoli indefiniti in una lingua specifica. Gli articoli indefiniti, come "un" e "una" in italiano, introducono nomi indefiniti e generalmente indicano una mancanza di specificità nell'oggetto o nel concetto a cui si riferiscono. Gli articoli spesso sono obbligatori in certi contesti che richiedono riferimenti definiti o specifici. Gli articoli definiti o specifici possono anche grammaticalizzarsi da pronomi che significano 'un certo'. Poiché gli articoli accompagnano necessariamente i sostantivi, è in questi casi che è necessario indagare se il marcatore può occorrere in modo indipendente o meno; se può, è più probabile che sia un pronome e non un articolo. Gli articoli possono essere omofoni con pronomi e dimostrativi, e anche con classificatori.

Grambank assegna un valore “1” se i sostantivi indefiniti sono comunemente accompagnati da un articolo indefinito/non-specifico oppure se ci sono tre o più esempi di un dimostrativo o numero per uno che si è grammaticalizzato in un articolo indefinito. Assegna un valore “0” se non ci sono articoli indefiniti o se ci sono meno di tre esempi di articoli indefiniti/non specifici che si verificano con sostantivi indefiniti. E infine un valore “?” se non è chiaro se ci siano articoli indefiniti a causa di un'analisi limitata da parte dell'autore o perché non ci sono abbastanza esempi. All'interno del database il valore “1” di Grambank è indicato come “presente” e il valore “0” come assente.

- **GB035-Are there three or more distance contrasts in demonstratives?**

Questo fattore cerca di determinare se una lingua specifica ha dimostrativi che distinguono tra almeno tre distanze diverse. Ad esempio, in alcune lingue, ci potrebbero essere dimostrativi per oggetti vicini, medi e lontani, e questa feature mira a evidenziare la presenza o l'assenza di tali distinzioni nella struttura grammaticale della lingua analizzata. I dimostrativi sono modificatori che possono alterare sostantivi per indicare la posizione (o un riferimento all'interno del discorso). Ciò non riguarda solo la distanza

euclidea (metri, pollici, ecc.), ma può anche essere relativo alla vicinanza o alla lontananza rispetto agli interlocutori.

Grambank assegna il valore “1” se ci sono tre o più dimostrativi adnominali che indicano gradi diversi di distanza, compresa la distanza dal parlante o dall'interlocutore. Assegna il valore “0” se ci sono meno di tre dimostrativi adnominali che indicano gradi diversi di distanza, compresa la distanza dal parlante o dall'interlocutore. All'interno del database il valore “1” di Grambank è indicato come “presente” e il valore “0” come assente.

- **GB036-Do demonstratives show an elevation distinction?**

Questo fattore cerca di determinare se i dimostrativi di una lingua specifica differenziano tra oggetti situati in posizioni più alte o più basse rispetto a un punto di riferimento. L'elevazione riguarda quanto un oggetto sia alto o basso (in altitudine). Ad esempio, alcune lingue possono avere dimostrativi che indicano non solo la distanza orizzontale, ma anche se un oggetto è in alto o in basso rispetto all'oratore o al destinatario.

Grambank trova i dimostrativi adnominali e assegna il valore “1” se ci sono dimostrativi adnominali che comunicano specificamente una distinzione di elevazione al parlante o all'interlocutore, piuttosto che solo indicare la vicinanza o la lontananza come in GB035. All'interno del database il valore “1” di Grambank è indicato come “presente” e il valore “0” come assente.

- **GB037-Do demonstratives show a visible-nonvisible distinction?**

Il procedimento attuato in Grambank consiste nel verificare se i dimostrativi mostrano una distinzione tra "visibile" e "non visibile" quando si fa riferimento a oggetti che si trovano alla stessa distanza relativa. Ad esempio, alcuni sistemi di dimostrativi potrebbero includere forme diverse per indicare oggetti che sono a portata di vista e oggetti che non sono visibili o sono nascosti.

Grambank assegna il valore “1” se ci sono dimostrativi adnominali che comunicano specificamente che qualcosa è invisibile/visibile per il parlante o l'interlocutore, anziché semplicemente indicare la distanza vicina o lontana, altrimenti

assegna il valore “0”. All’interno del database il valore “1” di Grambank è indicato come “presente” e il valore “0” come assente.

- **GB522-Can the S or A argument be omitted from a pragmatically unmarked clause when the referent is inferrable from context ('pro-drop' or 'null anaphora')?**

Questo fattore si riferisce alla possibilità di omettere il soggetto (S) o l'agente (A) in una proposizione pragmaticamente non marcata quando il referente può essere dedotto dal contesto. Questo fenomeno è noto come "pro-drop" o "null anaphora". In linguistica, alcune lingue consentono la caduta di pronomi o argomenti quando il significato può essere dedotto chiaramente dal contesto circostante. Questa feature si occupa di stabilire se, in una lingua specifica, è possibile omettere il soggetto o l'agente in una proposizione senza che ciò comprometta la chiarezza o la comprensione, purché il riferimento sia inferibile dal contesto.

Grambank assegna un valore “1” Se l'argomento S o A può essere omesso in una proposizione principale pragmaticamente non marcata quando il suo referente può essere dedotto dal contesto e un valore “0” se l'argomento S o A è sempre espresso fonologicamente da un sintagma nominale o da un pronome fonologicamente indipendente in una proposizione principale pragmaticamente non marcata. All’interno del database il valore “1” di Grambank è indicato come “presente” e il valore “0” come assente.

Dal WALs sono stati considerati:

- **107A- *Passive Constructions***

Nel WALs Anna Siewierska (2013) classifica una costruzione come passiva se presenta le seguenti cinque proprietà:

1. si distingue da un'altra costruzione, l'attiva;
2. il soggetto dell'attiva può corrispondere a una frase obliqua opzionale nella passiva o può non essere esplicitamente espresso;
3. il soggetto della passiva, se presente, è collegato all'oggetto diretto dell'attiva;
4. la costruzione è limitata dal punto di vista pragmatico rispetto all'attiva;

5. la costruzione presenta una marcatura morfologica speciale sul verbo.

Le costruzioni passive si verificano nel 44% delle lingue del campione del WALS. Sono più comuni tra le lingue dell'Eurasia e dell'Africa. Si trovano regolarmente anche nelle Americhe. Sono un po' meno frequenti nel sud-est asiatico e nel Pacifico. In Nuova Guinea sembra che non si attestino affatto. In Eurasia i passivi sono frequenti ovunque tranne che nel Caucaso e nelle lingue tibetane dell'India e del Nepal. In Africa i passivi sono molto comuni tra le lingue nilo-sahariane. Delle lingue Niger-Congo nel campione solo circa la metà mostra costruzioni passive. I soggetti passivi sono meno frequenti lungo le coste dell'Africa occidentale (Anna Siewierska, 2013).

Dopo aver analizzato la presenza o meno di questi fenomeni all'interno dei tre campioni, si è proceduto con un'analisi statistica utilizzando il test esatto di Fischer, allo scopo di valutare se i dati dicotomici provenienti dai due campioni, riassunti in una tabella di contingenza 2x2, siano in accordo con l'ipotesi nulla (H_0) che le popolazioni di origine dei due campioni presentino la stessa suddivisione dicotomica, e che le divergenze osservate nei dati campionari siano attribuibili esclusivamente al caso.

4. Risultati

La fase determinante di questa indagine è stata la formazione dei due campioni, con l'intento di rappresentare sia le lingue esclusivamente orali che quelle che presentano anche una tradizione scritta. I campioni di lingue sono stati creati in base al campione di 200 lingue del WALS, seguendo criteri quali la diffusione geografica, l'appartenenza a specifiche famiglie linguistiche e la quantità di testi disponibili. Questo approccio ha mirato a garantire una rappresentatività adeguata. Successivamente, è stata verificata la distribuzione di parametri specifici all'interno dei campioni al fine di facilitare una comprensione approfondita delle caratteristiche linguistiche nelle due modalità comunicative.

Il campione di lingue prive di tradizione scritta conta 176 lingue. Le restanti 24 lingue originariamente presenti nel campione del WALS non sono state sostituite. Dall'altro lato, il gruppo di lingue con tradizione scritta è composto da 147 lingue, di cui 53 non sono state sostituite rispetto al campione iniziale del WALS (tabella S1-S2-S3 in Appendice). Osservando le tabelle indicate, la differenza numerica dei due campioni evidenzia chiaramente che il campione di lingue con una tradizione scritta è inferiore rispetto a quello di lingue esclusivamente orali. Questa circostanza sottolinea come la mancanza di dati esaustivi possa ostacolare la comprensione completa della storia e della struttura linguistica di numerose lingue all'interno del campione.

I risultati derivanti dalle analisi delle famiglie linguistiche all'interno di ciascun campione sono i seguenti:

- Nel campione iniziale del WALS, sono incluse 103 famiglie linguistiche. Come può essere osservato dal grafico a torta (figura 4), la percentuale più significativa di lingue comprese in questo campione appartiene alle famiglie linguistiche Austronesiana ($15/200=14\%$), Niger-Congo ($13/200=12\%$) e Indoeuropea ($11/200=10\%$). Le categorie denominate "Altre" comprendono famiglie linguistiche caratterizzate dalla presenza di solamente una o due lingue (sono presenti 156 famiglie con all'interno una lingua, 15 con due lingue).

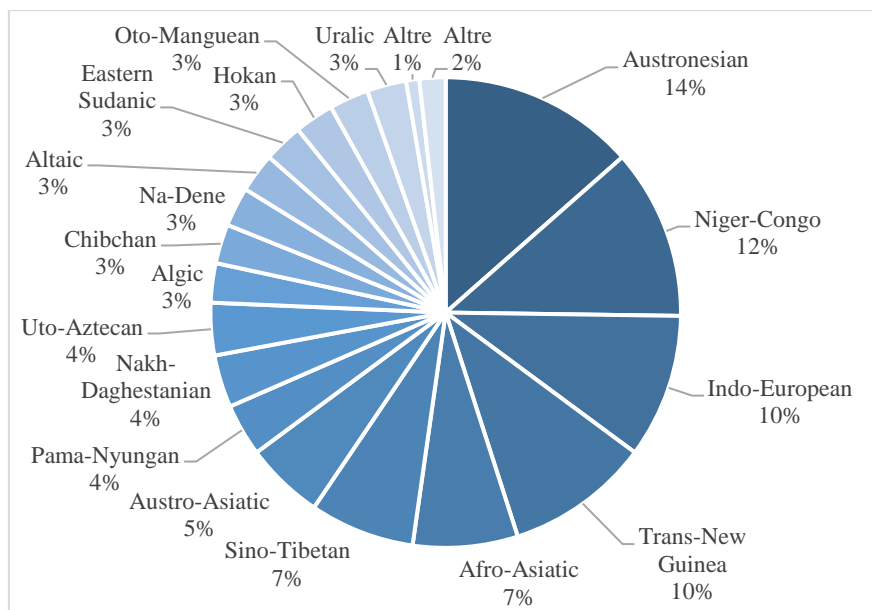


Figura 4: Campione WAL

- Nel campione delle lingue esclusivamente orali sono rappresentate 88 famiglie linguistiche. In questo campione la maggioranza delle lingue proviene dalla famiglia Niger-Congo (15:176=16%), seguita dalle famiglie Austronesiana e Trans-New Guinea, che comprendono entrambe 12 lingue (12:176=13%). Analogamente al caso precedente, le categorie denominate "Altre" incorporano famiglie linguistiche caratterizzate dalla presenza di esclusivamente una o due lingue, contando 64 famiglie con una sola lingua rappresentata e 11 con due lingue (figura 5).

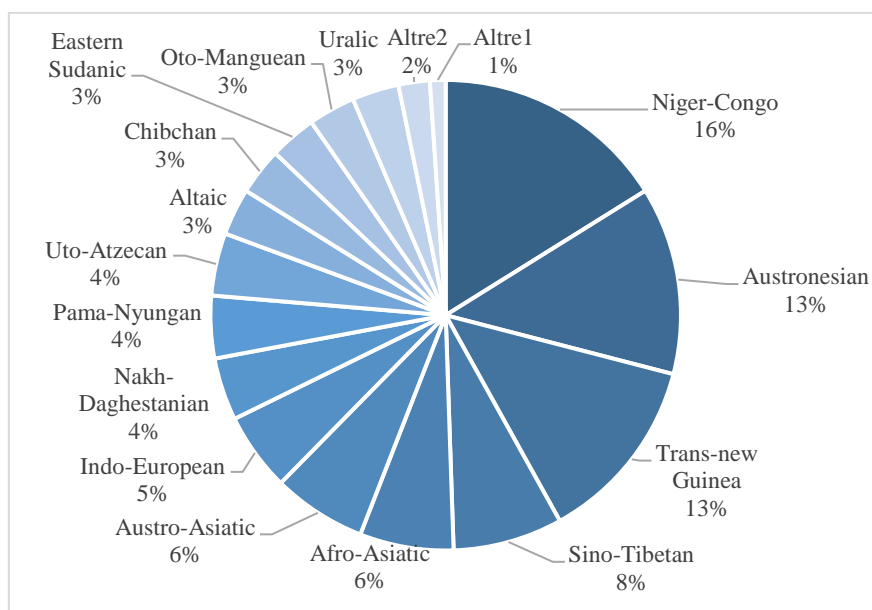


Figura 5: lingue senza tradizione scritta

- Nel campione delle lingue con tradizione scritta sono state identificate 66 famiglie linguistiche. In linea con il campione WALS, le famiglie linguistiche maggiormente rappresentate sono l'Austronesiana, l'Indoeuropea e la Niger-Congo, con un totale di 13 (13:147=14%) e 11 (11:147=12%) lingue rispettivamente. Le famiglie classificate come "altre" sono costituite da 50 famiglie con una sola lingua e 7 famiglie con due lingue (figura 6).

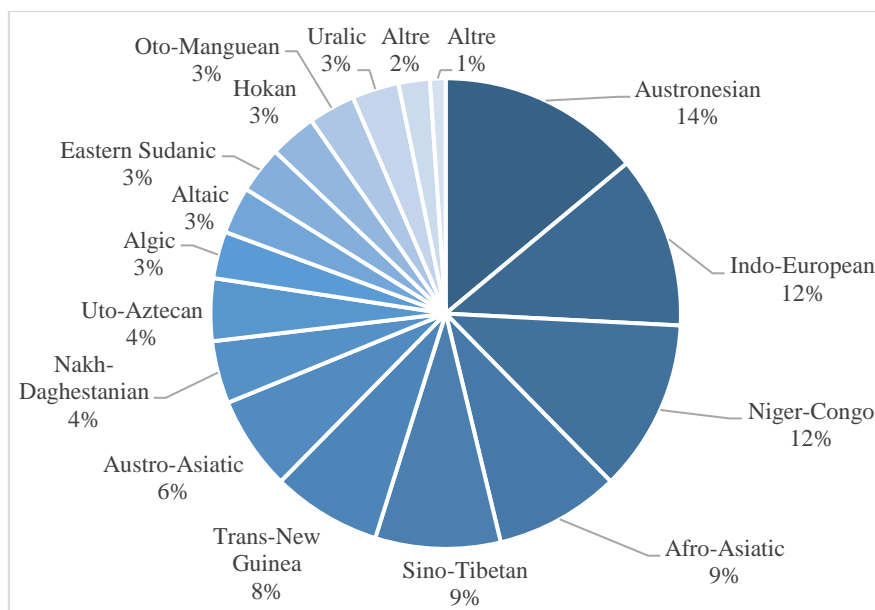


Figura 6: lingue con tradizione scritta

L'analisi del database sviluppato utilizzando i dati di Ethnologue ha fornito i seguenti risultati:

- La divisione delle lingue identificate nel campione di lingue esclusivamente orali e in quello di lingue scritte risulta essere accurata.
- Il campione iniziale di 200 lingue del WALS presenta 101 lingue esclusivamente orali (50,5 %), 71 lingue (35,5 %) che manifestano anche una tradizione scritta, mentre 28 lingue (14 %) sono categorizzate come "don't know" a causa di una disponibilità insufficiente di informazioni per una classificazione definitiva (come precedentemente illustrato). Questo indica una prevalenza significativamente superiore di lingue esclusivamente orali nel campione del WALS.

In merito all'esame della distribuzione dei fenomeni all'interno campioni, sono stati conseguiti i risultati seguenti:

- **GB020-*Are there definite or specific articles?***

Come si evince dai dati in tabella, nel campione di lingue orali, la presenza di lingue che richiedono l'uso di articoli definiti o specifici è inferiore rispetto a quelle che non lo richiedono (11,4% contro 22,2%). Ciò potrebbe indicare che nelle conversazioni verbali informali o colloquiali, come sono quelle solitamente elicitate nella descrizione di una lingua, l'uso di articoli specifici è meno comune. Nel campione di lingue scritte, la percentuale di lingue che presentano articoli definiti o specifici è, invece, più elevata rispetto al campione di lingue orali (24,4%), ma comunque inferiore alla percentuale di lingue scritte in cui gli articoli definiti o specifici non occorrono (46%). Questo potrebbe suggerire che la forma scritta richieda un uso più frequente di articoli specifici. Nelle lingue scritte, come l'italiano, è infatti comune utilizzare gli articoli indefiniti come "un" o "una" per introdurre un sostantivo non specifico o non precedentemente menzionato, spesso omessi più frequentemente in contesti informali o colloquiali. Nel campione misto del WALS, la presenza di lingue con articoli definiti o specifici è più elevata rispetto agli altri due campioni. Nella terza colonna le percentuali indicano le lingue per cui non si hanno tali dati. Come si può osservare, la percentuale di lingue orali prive di informazioni risulta essere superiore rispetto a quella rilevata nei restanti due campioni.

	PRESENT	ABSENT	NO DATA
ORAL	11,4%	22,2%	66,4%
WRITTEN	24,4%	46%	29,6%
MIX (WALS)	27%	50%	23%

Tabella 1: frequenza della presenza o assenza di articoli definiti nei tre campioni

Il test esatto di Fisher evidenzia che c'è una probabilità di 1 di riscontrare tali frequenze (Tabella 2). Di conseguenza, il risultato non raggiunge significatività statistica con un valore di $p > 0,05$ e non c'è sufficiente supporto statistico per affermare che vi sia un'associazione significativa tra le variabili in esame.

	PRESENT	ABSENT	MARGINAL TOTALS	ROW
ORAL	20	39	59	
WRITTEN	34	64	98	
MARGINAL	54	103	157 (Grand Total)	
COLUMN TOTALS				

Tabella 2: test Fisher-articoli definiti

- **GB021-Do indefinite nominals commonly have indefinite articles?**

Analizzando la tabella 3, si evince che nel campione di lingue esclusivamente orali, la percentuale di lingue in cui i nomi indefiniti richiedono l'uso di articoli indefiniti è molto bassa (4:175=2,2%), mentre la percentuale di lingue che richiedono l'uso di nomi indefiniti senza articoli è notevolmente più alta (29,7%). Nelle lingue con tradizione scritta, la percentuale di lingue in cui i nomi indefiniti richiedono l'uso di articoli indefiniti è più elevata (17:139=12,2%), ma ancora inferiore rispetto a quelle in cui i nomi indefiniti occorrono senza articoli (53,9%). Questo indica che, sia nelle lingue scritte che in quelle orali, gli articoli indefiniti con nomi indefiniti non sono così comuni. Nel campione misto del WALS, la percentuale di nomi indefiniti con articoli indefiniti è dell'11,5%, quindi ancora inferiore rispetto a quelli senza articoli. In generale, in tutti e tre i campioni, la presenza di articoli indefiniti con nomi indefiniti è inferiore rispetto alla loro assenza. Nella terza colonna le percentuali indicano le lingue per cui non si hanno tali dati. Nella terza colonna le percentuali indicano le lingue per cui non si hanno tali dati. Come si può osservare, la percentuale di lingue orali prive di informazioni risulta essere superiore rispetto a quella rilevata nei restanti due campioni.

	PRESENT	ABSENT	NO DATA
ORAL	2,2%	29,7%	68,1%
WRITTEN	12,2%	53,9%	33,9%
MIX (WALS)	11,5%	65%	23,5%

Tabella 3: frequenza della presenza o assenza di nomi indefiniti con articoli indefiniti nei tre campioni

Il test esatto di Fisher mostra che c'è una probabilità di 0,0317 di osservare queste frequenze (tabella 4). Il valore p ottenuto dal test di Fisher è 0,10, come descritto nella tabella. Il risultato è significativo a $p < 0,10$. Quindi, c'è una probabilità del 10% che i risultati osservati siano dovuti al caso, ma potrebbe indicare una tendenza verso la significatività.

	PRESENT	ABSENT	MARGINAL ROW TOTALS
ORAL	4	52	56
WRITTEN	17	65	82
MARGINAL	21	117	138(Grand Total)
COLUMN TOTAL			

Tabella 4- Test Fisher: nomi indefiniti con articoli indefiniti

- **GB035-*Are there three or more distance contrasts in demonstratives?***

La tabella 5 indica che la percentuale di lingue orali che presentano dimostrativi con tre o più contrasti di distanza è marginalmente superiore rispetto a quella delle lingue orali che ne sono sprovviste con una percentuale del 16% rispetto a 14,8%. Anche la percentuale di lingue con tradizione scritta che presentano dimostrativi con tre o più contrasti di distanza è leggermente più elevata rispetto a alla percentuale di lingue che ne sono prive. Nel campione misto del WALS, la percentuale di lingue che richiedono l'uso di dimostrativi con tre o più contrasti di distanza è ancora più alta rispetto agli altri due campioni con una percentuale del 40,5%. Sembra che ci sia una tendenza verso una maggiore presenza di lingue che presentano dimostrativi con tre o più contrasti di distanza rispetto alla loro assenza in tutti e tre i campioni. Tuttavia, è importante notare che le differenze tra le percentuali non sono molto marcate, e le variazioni possono essere influenzate da diversi fattori linguistici. Nella terza colonna le percentuali indicano le lingue per cui non si hanno tali dati. Come si può osservare, la percentuale di lingue orali prive di informazioni risulta essere superiore rispetto a quella rilevata nei restanti due campioni.

	PRESENT	ABSENT	NO DATA
ORAL	16%	14,8%	69,2%
WRITTEN	35,9%	34,5%	29,6%
MIX (WALS)	40,5%	37,5%	22%

Tabella 5: frequenza di presenza o assenza dei dimostrativi con tre o più contrasti di distanza all'interno dei tre campioni

Il test esatto di Fisher mostra che c'è una probabilità di 1 di osservare queste frequenze (tabella 6). Il risultato, quindi, non è significativo a $p < 0,05$. Non c'è sufficiente supporto statistico per affermare che vi sia un'associazione significativa tra le variabili in esame, la variabile "PRESENT/ABSENT" e la variabile "ORAL/WRITTEN".

	PRESENT	ABSENT	MARGINAL	ROW
ORAL	28	26	54	
WRITTEN	50	48	98	
MARGINAL	78	74	152 (Grand Total)	
COLUMN TOTALS				

Tabella 6: Test Fisher- dimostrativi con tre o più contrasti di distanza

- **GB036-Do demonstratives show an elevation distinction?**

Come evidenziato nella tabella 7, nelle lingue senza tradizione scritta, la percentuale di lingue che presentano dimostrativi con distinzione di elevazione è piuttosto bassa (3,4%), mentre la percentuale di lingue in cui i dimostrativi omettono questa distinzione è notevolmente più alta (28%). Ciò suggerisce che, nel linguaggio parlato, l'uso di dimostrativi con una distinzione di elevazione sia meno comune. Nelle lingue con tradizione scritta, la percentuale di lingue che richiedono l'uso di dimostrativi con distinzione di elevazione è leggermente superiore rispetto alle lingue orali. Questo indica un uso più frequente di queste forme nei contesti scritti. Tuttavia, la presenza è ancora relativamente bassa rispetto all'assenza. Nel campione misto del WALS, la percentuale di lingue che presentano l'uso di dimostrativi con distinzione di elevazione è più alta rispetto agli altri due campioni. Anche qui, l'assenza di questa distinzione è significativamente più comune. In generale, sembra che la distinzione di elevazione nei dimostrativi sia meno comune rispetto all'assenza di tale distinzione in tutti e tre i campioni. Tuttavia, va notato che la distinzione è particolarmente bassa nel campione di lingue orali, mentre negli altri due campioni è leggermente più frequente. Nella terza colonna le percentuali indicano le lingue per cui non si hanno tali dati. Come si può osservare, la percentuale di lingue orali prive di informazioni risulta essere superiore rispetto a quella rilevata nei restanti due campioni.

	PRESENT	ABSENT	NO DATA
ORAL	3,4%	28%	68,6%
WRITTEN	5%	64,7%	30,3%
MIX (WALS)	8%	70%	22%

Tabella 7: frequenza della presenza o assenza di dimostrativi con distinzione di elevazione nei tre campioni

Il test esatto di Fisher mostra che c'è una probabilità di 0,5479 di osservare queste frequenze (tabella 8). Il risultato, quindi, non è significativo a $p < 0,05$. Non c'è un'associazione significativa tra la variabile "PRESENT/ABSENT" e la variabile "ORAL/WRITTEN".

	PRESENT	ABSENT	MARGINAL	ROW
ORAL	6	49	55	
WRITTEN	7	90	97	
MARGINAL	13	139	152	
COLUMN TOTALS				

Tabella 8: Test Fisher- dimostrativi con distinzione di elevazione

- **GB037-Do demonstratives show a visible-nonvisible distinction?**

Osservando la tabella 9, si evince che nel campione di lingue senza tradizione scritta, la percentuale di lingue che presentano dimostrativi che mostrano una distinzione tra visibile e non visibile è piuttosto bassa (4%), mentre la percentuale di lingue in cui i dimostrativi occorrono senza questa distinzione è notevolmente più alta (26,8%). Ciò suggerisce che, nel linguaggio parlato, l'uso di dimostrativi con questa distinzione è meno comune. Nel campione di lingue scritte, la percentuale di lingue che richiedono l'uso di dimostrativi con distinzione tra visibile e non visibile è leggermente superiore rispetto a quelle senza questa distinzione, indicando un uso più frequente di queste forme nei contesti scritti. Tuttavia, anche qui, la presenza è ancora relativamente bassa rispetto all'assenza. Nel campione misto del WALS, la percentuale di lingue che presentano dimostrativi con distinzione tra visibile e non visibile è più alta rispetto agli altri due campioni ma, anche in questo caso, l'assenza di questa distinzione è significativamente più comune. In generale, sembra che la distinzione tra visibile e non visibile nei dimostrativi sia meno comune rispetto all'assenza di tale distinzione in tutti e tre i campioni. Nella terza colonna le percentuali indicano le lingue per cui non si hanno tali dati. La percentuale di lingue orali prive di informazioni risulta essere superiore rispetto a quella rilevata nei restanti due campioni.

	PRESENT	ABSENT	NO DATA
ORAL	4%	26,8%	69,2%
WRITTEN	7,9%	60,4%	31,7%
MIX (WALS)	10%	66%	24%

Tabella 9: frequenza della presenza o assenza di dimostrativi che mostrano una distinzione tra visibile e non visibile nei tre campioni

Il test esatto di Fisher mostra che c'è una probabilità di 0,79 di osservare queste frequenze (tabella 10). Il risultato, quindi, non è significativo a $p < 0,05$. Non c'è sufficiente supporto statistico per affermare che vi sia un'associazione significativa tra le variabili in esame.

	PRESENT	ABSENT	MARGINAL TOTAL	ROW
ORAL	7	47	54	
WRITTEN	11	84	95	
MARGINAL	18	131	149 (Grand Total)	
COLOUMN TOTALS				

Tabella 10: Test Fisher- dimostrativi che mostrano una distinzione tra visibile e non visibile

- **GB522-Can the S or A argument be omitted from a pragmatically unmarked clause when the referent is inferrable from context ("pro-drop" or "null anaphora")?**

Nel campione di lingue esclusivamente orali, la percentuale di lingue in cui è possibile omettere l'argomento S o A quando il referente è inferibile dal contesto è relativamente alta (19,4%), mentre la percentuale di lingue in cui questo non è possibile è molto bassa (3,4%). Ciò suggerisce che nel linguaggio parlato c'è una tendenza a usare l'omissione del soggetto. Nelle lingue scritte, la percentuale in cui è possibile omettere l'argomento S o A è, però, ancora più elevata (33,8%), ma l'omissione è comunque maggiore rispetto alle lingue orali (10,7%). Questo potrebbe indicare che anche nei contesti scritti l'omissione degli argomenti sia utilizzata, ma in misura minore rispetto alle lingue orali. Per quanto riguarda il campione misto del WALS, la tendenza al pro drop è ancora più accentuata. La percentuale di lingue in cui è possibile omettere l'argomento S o A è alta con un valore pari al 37%, mentre quella in cui questo non è possibile è relativamente bassa con un valore dell'11,5% (tabella 11). Nella terza colonna le percentuali indicano le lingue per cui non si hanno tali dati. Come si può osservare, la percentuale di lingue orali prive di informazioni risulta essere superiore rispetto a quella rilevata nei restanti due campioni.

	PRESENT	ABSENT	NO DATA
ORAL	19,4%	3,4%	77,2%
WRITTEN	33,8%	10,7%	55,5%
MIX (WALS)	37%	11,5%	51,5%

Tabella 11: frequenza della presenza o assenza di pro drop nei tre campioni

Il test esatto di Fisher mostra che c'è una probabilità di 0,3215 di osservare queste frequenze (tabella 12). Il risultato, quindi, non è significativo a $p < 0,05$. Non c'è un'associazione significativa tra la variabile "PRESENT/ABSENT" e la variabile "ORAL/WRITTEN".

	PRESENT	ABSENT	MARGINAL TOTAL	ROW
ORAL	34	6	40	
WRITTEN	47	15	62	
MARGINAL	81	21	102 (Grand Total)	
COLOUMN TOTALS				

Tabella 11: Test Fisher-pro drop

- **107A- Passive Constructions**

Analizzando la tabella 13, si evince che il numero di lingue con tradizione scritta che presentano costruzioni passive è maggiore rispetto a quello delle lingue esclusivamente orali. Il 29,5% delle lingue che hanno una tradizione scritta utilizzano costruzioni passive, mentre solo l' 8% di quelle parlate adotta la stessa struttura. Nel campione misto delle lingue del WALs, la presenza delle lingue con questa caratteristica è ancora più alta, registrando il 40%, rispetto al 51% in cui è assente. Questo suggerisce che, all'interno del campione misto la caratteristica è più prevalente. Nella terza colonna le percentuali indicano le lingue per cui non si hanno tali dati. Come si può osservare, la percentuale di lingue orali prive di informazioni risulta essere superiore rispetto a quella rilevata nei restanti due campioni.

	PRESENT	ABSENT	NO DATA
ORAL	8%	15%	77%
WRITTEN	29,4%	23%	47,6%
MIX (WALS)	40%	51%	9%

Tabella 12: frequenza della presenza o assenza di costruzioni passive nei tre campioni

Il test esatto di Fisher mostra che c'è una probabilità di 0,0317 di osservare queste frequenze. La relazione tra queste variabili è risultata essere statisticamente significativa con un valore p di 0,05, come indicato nella tabella 14. Ciò implica che c'è una probabilità inferiore al 5% che la relazione osservata tra lingue parlate e scritte e l'uso di costruzioni passive sia dovuta al caso. In altre parole, i risultati suggeriscono che vi è un legame significativo tra il tipo di lingua (orale o scritta) e la frequenza delle costruzioni passive, supportando l'ipotesi iniziale.

	PRESENT	ABSENT	MARGINAL TOTAL	ROW
ORAL	14	27	41	
WRITTEN	41	32	73	
MARGINAL	55	59	114 (Grand Total)	
COLOUMN TOTALS				

Tabella 13: Test Fisher-costruzioni passive

Le possibili ragioni per cui le lingue scritte spesso adottano più frequentemente costruzioni passive sono diverse. Una di queste è la capacità delle costruzioni passive di conferire al testo un tono più formale e obiettivo. Questo avviene attraverso la diminuzione dell'accento sull'agente (chi compie l'azione), rendendo la lingua più neutrale

e impersonale. In aggiunta, in situazioni in cui l'identità dell'agente non è nota, l'utilizzo di una costruzione passiva permette di mantenere l'attenzione sull'azione stessa. La frequente adozione di tali costruzioni può anche essere influenzata dalle convenzioni linguistiche consolidate nella storia della lingua.

5. Discussioni e conclusioni

La vastità della diversità linguistica mondiale, come evidenziato dai dati forniti da Ethnologue, mette in luce la complessità e la ricchezza delle lingue esistenti. Tuttavia, emerge una preoccupante realtà: molte di queste lingue, che sono spesso lingue esclusivamente orali, sono minacciate o in via di estinzione (Krauss 1992; Miyaoka et al. 2007; Austin e Sallabank 2011, Nettle 2000, Crystal 2002). L'impegno nella conservazione e documentazione di queste lingue diventa cruciale per preservare non solo la lingua in sé, ma anche la ricchezza culturale e antropologica associata. Il numero significativo di lingue prive di tradizione scritta e senza documentazione rappresenta una sfida inestimabile per la conservazione del patrimonio linguistico globale (Adda et al., 2016). In questo contesto, l'obiettivo della presente indagine è stato quello di investigare le disparità e le peculiarità tra le lingue che utilizzano la forma scritta e quelle che si basano sulla forma orale. Ciò è stato conseguito attraverso la formulazione di due campioni linguistici e l'analisi della distribuzione di specifici fenomeni linguisticamente rilevanti. La creazione dei campioni ha sottolineato un dato noto: le lingue orali sono più diffuse a livello globale rispetto a quelle scritte (Goody, 1986). In effetti, il campione delle lingue orali creato presenta un numero superiore rispetto a quello delle lingue che presentano anche una forma scritta. Tuttavia, va notato che, nonostante la loro maggiore diffusione, le lingue esclusivamente orali tendono a ricevere meno attenzione e a essere meno studiate rispetto alle lingue scritte, con conseguente minore disponibilità di informazioni su di esse. Lo studio ha, quindi, chiaramente evidenziato la persistenza del bias della lingua scritta già identificato da Linell (1982, 2004, 2012). Questo conferma una carenza di documentazione per le lingue orali, sottolineando la necessità di riconsiderare il modo in cui ci si approccia e si valorizzano queste ultime. I motivi di questa carenza, come descritto da Grandi & Mauri (2022), sono:

- Limitata accessibilità e diversità socioculturale e geografico-politica delle comunità: lingue parlate da comunità ristrette o isolate possono essere meno accessibili per gli studiosi. La mancanza di risorse, infrastrutture e contatti diretti con le comunità linguistiche può rendere difficile lo studio di tali lingue.

- Mancanza di risorse: le lingue senza tradizione scritta spesso mancano di risorse linguistiche come dizionari, grammatiche e testi scritti. Questa mancanza di documentazione rende più difficile per gli studiosi acquisire una comprensione approfondita di queste lingue.
- Risorse finanziarie: Gli studi linguistici richiedono risorse finanziarie per la ricerca sul campo, la registrazione audio, la documentazione e la formazione di ricercatori.
- Priorità di ricerca: Le lingue più diffuse o quelle con una ricca tradizione scritta possono essere prioritarie per gli studiosi, poiché la loro documentazione e analisi possono avere un impatto più ampio sulla comunità scientifica e sulla società.

I risultati di questa indagine, inoltre, si allineano con le teorie linguistiche (Antonelli, 2009; Voghera, 2017; Linell, 1982; Auer, 2009; Farrel, 1985) che evidenziano le divergenze tra le lingue scritte e orali. L'analisi delle tabelle dalla 1 alla 14 rivela che le lingue scritte sono caratterizzate da una maggiore complessità sintattica e lessicale rispetto alle lingue orali esaminate e mostrano una tendenza a utilizzare strutture grammaticali più elaborate con molte subordinate e costruzioni passive, suggerendo una maggiore attenzione alla precisione, alla formalità e allo stile. Inoltre, la lingua scritta è più stabile e permanente, conservando le informazioni nel tempo e gioca un ruolo cruciale nella cultura e nella comunicazione. D'altra parte, le lingue orali presentano una maggiore varietà di espressioni colloquiali e una sintassi più flessibile, evidenziando la natura dinamica e contestuale della comunicazione parlata.

Il bias della lingua scritta, però, ha influito sui risultati ottenuti per quanto riguarda l'analisi della distribuzione di alcuni fenomeni. Al fine di ottenere una comprensione più completa e accurata, sarebbe essenziale esplorare ulteriormente i fenomeni analizzati e ottenere più dati per confermare o confutare le conclusioni emerse. Ricerche fondamentali come quella condotta da Adda et al. (2016), mirate a colmare il vuoto tecnologico attuale per una documentazione efficiente delle lingue non scritte, assumono un'importanza cruciale. Secondo Linell, infatti, se si vuole comprendere la vera natura del linguaggio, è necessario considerare la parola e la scrittura come due media diversi, ciascuno con le proprie caratteristiche che vale la pena indagare (Linell, 2004). Allo stesso modo, l'indagine di Boito et al. (2017), focalizzata sulle lingue in via di estinzione e non scritte parlate in piccole comunità globali, offre proposte concretamente pratiche, come la pianificazione di attività e la suggestione di canali per l'estrazione automatica di

informazioni dai segnali vocali. L'impegno dell'UNESCO (si veda: <https://www.unesco.org/en>) e del SIL International (si veda: <https://www.sil.org/language-programs-management>) nel promuovere lo studio linguistico delle lingue non scritte, mirato alla produzione di materiali linguisticamente validi per i programmi di alfabetizzazione nelle lingue native, sottolinea ulteriormente l'importanza della documentazione delle lingue orali. Crowley (2007), nel sostenere che la documentazione delle lingue orali sia di estrema importanza nel campo della linguistica, evidenzia che rispondere alle domande sulla natura dei linguaggi umani richiede un'attenzione particolare anche alle lingue meno diffuse. Studiare solo le grandi lingue scritte significherebbe trascurare la ricca diversità linguistica umana. Inoltre, con la documentazione limitata delle molte lingue più piccole del mondo, spesso ci troviamo impossibilitati a rispondere in modo definitivo a domande cruciali sulla funzionalità generale delle lingue, come sottolineato da Adams (2017). Studiando le lingue orali, anche Farrel (1985) sottolinea l'importanza di comprendere come queste influenzino la nostra concezione del tempo, dello spazio e delle relazioni umane.

Quindi, in linea con le ricerche precedentemente menzionate, comprendere meglio le peculiarità delle lingue orali e delle lingue scritte può contribuire a vari aspetti importanti:

- **Preservare e Valorizzare la Cultura Orale:**

Un'analisi approfondita delle lingue orali aiuta a preservare e valorizzare la cultura orale, inclusi miti, tradizioni, e storie tramandate oralmente. Ciò contribuisce alla conservazione di patrimoni culturali unici e all'apprezzamento della diversità culturale.

- **Sviluppare Programmi Educativi Efficaci:**

Comprendere le peculiarità delle lingue orali è essenziale per la progettazione di programmi educativi efficaci. Tale consapevolezza permette di sviluppare strategie di insegnamento che tengono conto delle caratteristiche specifiche della trasmissione orale del sapere.

- **Preservare le Lingue in Pericolo di Estinzione:**

Studiare le peculiarità delle lingue orali può contribuire agli sforzi di preservare le lingue in pericolo di estinzione. Capire come queste lingue vengono trasmesse oralmente

fornisce informazioni preziose per sviluppare strategie di revitalizzazione e mantenere vive queste comunità linguistiche.

- Promuovere la Diversità Linguistica:

Comprendere le differenze tra lingue scritte e orali sottolinea l'importanza della diversità linguistica. Questa consapevolezza può promuovere l'apprezzamento delle molteplici espressioni linguistiche presenti nelle varie comunità globali.

- Favorire la Ricerca Etnografica e Antropologica:

Per gli studiosi di etnografia e antropologia, la comprensione delle lingue orali e scritte è fondamentale per condurre ricerche accurate e rispettose delle pratiche culturali e sociali di specifiche comunità.

La conservazione e la documentazione delle lingue orali prive di tradizione scritta non sono solo imperativi necessari per preservare la diversità linguistica, ma rappresentano altresì un contributo inestimabile nel campo della linguistica in generale. Questa importanza assume particolare rilevanza considerando l'ampia diffusione numerica delle lingue orali, che rappresentano una parte significativa del patrimonio linguistico mondiale. La mancanza di informazioni dettagliate e il bias derivante dalla focalizzazione sulla lingua scritta emergono come ostacoli significativi per una comprensione approfondita della storia e della struttura linguistica della maggior parte delle lingue del mondo. Le lingue orali, con le loro caratteristiche peculiari, si rivelano come un terreno fertile per ulteriori esplorazioni in ambito linguistico, aprendo la strada a studi più approfonditi su questo aspetto.

In conclusione, si può affermare che la presente indagine svolge un ruolo importante data la limitata quantità di studi paralleli. La scarsa presenza di ricerche simili rappresenta un punto di partenza per ulteriori futuri approfondimenti. Il risultato principale di questo lavoro è, quindi, quello di aver costituito dei campioni utilizzabili ed effettivamente funzionanti per verificare il bias della lingua scritta.

Bibliografia

- Adams, O. (2017). Automatic understanding of unwritten languages (Doctoral dissertation, Ph. D. thesis, University of Melbourne).
- Adda, G., Stüker, S., Adda-Decker, M., Ambourou, O., Besacier, L., Blachon, D., ... & Zerbian, S. (2016). Breaking the unwritten language barrier: The BULB project. *Procedia Computer Science*, 81, 8-14. <https://doi.org/10.1016/j.procs.2016.04.023>
- Antonelli, G., Motolese, M., & Tomasin, L. (2014). *Storia dell'italiano scritto* (Vol. 3, pp. 1-499). Carocci.
- Antonelli, G. (2019). Parlare, scrivere, digitare. In *L'italiano. Parlare, scrivere, digitare* (pp. 7-29). Treccani.
- Auer, P., Couper-Kuhlen, E., & Muller, F. (1999). *Language in time: The rhythm and tempo of spoken interaction*. Oxford University Press.
- Auer, P. (2009). On-line syntax: Thoughts on the temporality of spoken language. *Language Sciences*, 31(1), 1-13. <https://doi.org/10.1016/j.langsci.2007.10.004>
- Austin, P. K., & Sallabank, J. (Eds.). (2011). *The Cambridge handbook of endangered languages*. Cambridge University Press.
- Bell, A. (1978). Language samples. *Universals of human language*, ed. by Joseph Greenberg et al., 1.153-202.
- Bird, S., & Simons, G. F. (2022). The Open Language Archives Community: a 20-year update. *The Electronic Library*, 40(5), 507-524.
- Bloomfield, L. (1933). *Language*. -London: Allen & Unwin.
- Boito, M. Z., Bérard, A., Villavicencio, A., & Besacier, L. (2017, December). Unwritten languages demand attention too! word discovery with encoder-decoder models. In *2017 IEEE automatic speech recognition and understanding workshop (ASRU)* (pp. 458-465). IEEE.
- Chomsky, N. (1965). *Aspects of the Theory of Syntax* (No. 11). MIT press.

- Croft, W. (2002). *Typology and universals*. Cambridge University Press.
- Crystal, D. (2002). *Language death*. Cambridge University Press.
- Derrida, J. (1967). La linguistique de Rousseau. *Revue Internationale de philosophie*, 443-462. <https://www.jstor.org/stable/45104863>
- Dittmar, Norbert. Beschreibungen für gesprochenes Deutsch auf dem Prüfstand: Analysen und Perspektiven. Ed. Nils Bahlo. Peter Lang, 2010.
- Dryer, M. S. (1989). Large linguistic areas and language sampling. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language,"* 13(2), 257-292. <https://doi.org/10.1075/sl.13.2.03dry>
- Farrell, T. J. (1985). *Orality and Literacy: The Technologizing of the Word*.
- Georgi, R., Xia, F., & Lewis, W. (2010, August). Comparing language similarity across genetic and typologically based groupings. In *Proceedings of the 23rd international conference on computational linguistics (Coling 2010)* (pp. 385-393).
- Goodwin, C. (1979). The interactive construction of a sentence in natural conversation. *Everyday language: Studies in ethnomethodology*, 97, 101-121.
- Goodwin, C. (1980). Restarts, pauses, and the achievement of a state of mutual gaze at turn-beginning. *Sociological inquiry*, 50(3-4), 272-302.
- Goodwin, C. (1995). Sentence construction within interaction. *Aspects of oral communication*, 198-219.
- Goody, J. (1977). *The domestication of the savage mind*. Cambridge University Press.
- Goody, J. (1986). *The logic of writing and the organization of society*. Cambridge University Press.
- Goody, J. (2000). *The power of the written tradition*. (No Title).
- Grandi, N., & Mauri, C. (2022). *La tipologia linguistica. Unità e diversità nelle lingue del mondo* (pp. 1-343). Carocci. <https://hdl.handle.net/11585/893917>
- Hockett, C. F. (1958). *A course in modern linguistics*.

- Hockett, C. F., & Hockett, C. D. (1960). The origin of speech. *Scientific American*, 203(3), 88-97. <https://www.jstor.org/stable/24940617>
- Holman, E. W., Schulze, C., Stauffer, D., & Wichmann, S. (2007). On the relation between structural diversity and geographical distance among languages: observations and computer simulations. <https://doi.org/10.1515/LINGTY.2007.027>
<https://doi.org/10.1017/CBO9780511620874.005>
- Iannàccaro, G. (2000). Per una semantica più puntuale del concetto di dato linguistico: un tentativo di sistematizzazione epistemologica. *Per una semantica più puntuale del concetto di dato linguistico*, 1000-1029.
- Krauss, M. (1992). The world's languages in crisis. *Language*, 68(1), 4-10.
- Lehmann, C. (2004). *Data in linguistics*.
- Lepschy, G. C. (2002). *Mother tongues and other reflections on the Italian language*. University of Toronto Press.
- Lerner, G. H. (1996). On the "semi-permeable" character of grammatical units in conversation: Conditional entry into the turn space of another speaker. *Studies in interactional sociolinguistics*, 13, 238-276.
- Lerner, G. H. (2004, August). Collaborative turn sequences. In *Conversation*.
- Lewis, M. P., & Simons, G. F. (2010). *Assessing endangerment: expanding Fishman's GIDS*.
- Lewis, M. P. (2014). *Ethnologue: Languages of the world*. (No Title).
- Linell, P. (1982). The concept of phonological form and the activities of speech production and speech perception. *Journal of phonetics*, 10(1), 37-72.
- Linell, P. (2004). *The written language bias in linguistics: Its nature, origins, and transformations*. Routledge.
- Linell, P. (2012). On the nature of language: formal written-language-biased linguistics vs. dialogical language sciences. *Cognitive Dynamics in Linguistic Interactions*. Newcastle upon Tyne: Cambridge Scholars Publishing, 107-124.

- Linell, P. (2019). The Written Language Bias (WLB) in linguistics 40 years after. *Language Sciences*, 76, 101230. <https://doi.org/10.1016/j.langsci.2019.05.003>
- Lyons, J. (1981). *Language and linguistics*. Cambridge university press.
- Mana, N., Cosi, P., Tisato, G., Cavicchio, F., Magno, E. C., & Pianesi, F. (2006, May). An italian database of emotional speech and facial expressions. In *The Workshop Programme Corpora for Research on Emotion and Affect Tuesday 23 rd May 2006* (p. 68).
- Merritt, R. (1987). *A Guide to the World's Languages*.
- Miestamo, M., Bakker, D., & Arppe, A. (2016). Sampling for variety. *Linguistic Typology*, 20(2), 233-296. <https://doi.org/10.1515/lingty-2016-0006>
- Miyaoka, O. (2007). Linguistic diversity in decline: A functional view. *The Vanishing Languages of the Pacific Rim*, 144.
- Monica, B. (1994). Il parlato italiano contemporaneo, in Serianni, L./Trifone, P., eds., *Storia della lingua italiana, I*, Bari: Laterza. In No Title name available (pp. 239-270).
- Nencioni, G. (1976). Parlato-parlato, parlato-scritto, parlato-recitato Parlé-parlé, parlé-écrit, parlé-récité. *Strumenti critici*, 10(1), 1-56.
- Nettle, D. (1999). *Linguistic diversity*. Oxford University Press.
- Nettle, D., & Romaine, S. (2000). *Vanishing voices: The extinction of the world's languages*. Oxford: Oxford University Press.
- Nichols, J. (1992). (1992)-Linguistic diversity in space and time. Chicago, IL: University of Chicago Press Nichols Linguistic diversity in space and time 1992.
- Oxford Advanced Learner's Dictionary, 2010. Oxford University Press, Oxford, p. 8.
- Perkins, R. D. (1980). *The evolution of culture and grammar*. State University of New York at Buffalo.
- Perkins, R. D. (1988). The covariation of culture and grammar. *Studies in syntactic typology*, 359-378.
- Perkins, R. D. (1989). Statistical techniques for determining language sample size. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language,"* 13(2), 293-315.

- Rijkhoff, J., Bakker, D., Hengeveld, K., & Kahrel, P. (1993). A method of language sampling. *Studies in Language*. International Journal sponsored by the Foundation "Foundations of Language," 17(1), 169-203. <https://doi.org/10.1075/sl.17.1.07rij>
- Sapir, E. (1921). An introduction to the study of speech. *Language*, 1.
- Saussure, F. de. 1964. (1916). *Cours de linguistique generale*. Paris : Payot.
- Savy, Renata & Cutugno, Francesco (1999), Su alcune correlazioni tra riduzioni segmentali e tratti prosodici nel parlato spontaneo: il ruolo del fattore tempo, in *Atti del XXVII convegno nazionale dell'Associazione Italiana di Acustica* (Genova, 26-28 maggio 1999), Genova, Tipo-lito Sorriso Francescano, pp. 183-187.
- Schlegel, F. (1977). *Über die Sprache und Weisheit der Indier: Ein Beitrag zur Begründung der Altertumskunde* (Heidelberg, 1808) (Vol. 1). John Benjamins Publishing.
- Seiler, H. (2000). *Language universals research: A synthesis* (Vol. 8). Gunter Narr Verlag.
- Siewierska, A. (2005). Passive constructions. <https://doi.org/10.5281/zenodo.7385533>
- Skalička, V. (2013). *Typologische studien* (Vol. 11). Springer-Verlag.
- Sornicola, Rosanna (1981), *Sul parlato*, Bologna, il Mulino.
- Thormoset, H. J. (2006). *Learning to speak an unwritten language* (Doctoral dissertation, Univ. of Toronto Toronto).
- Turchetta, B. (2000). *La ricerca di campo in linguistica: metodi e tecniche d'indagine* (Vol. 151). Carocci Editore.
- Voghera, Miriam (1992), *Sintassi e intonazione nell'italiano parlato*, Bologna, il Mulino.
- Voghera, Miriam (2004), La distribuzione delle parti del discorso nel parlato e nello scritto, in *La variabilité en langue*, editors R. van Deyck, R. Sornicola & J. Kabatek, Gent, *Communication & Cognition*, 2 voll., vol. 1° (Langue parlée et langue écrite dans le présent et dans le passé), pp. 261-284.
- Voghera, M. (2005). La misura delle categorie sintattiche. In *Parole e numeri* (pp. 125-138). Aracne.

- Voghera, M. (2010). The Constants of Spoken Discourse: what evidence for grammatical models? In *Spoken communication between Symbolics and Deixis* (pp. 77-112). Cambridge Scholar Press.
- Voghera, M. (2013). Tipi di tipo nel parlato e nello scritto. *Tipi di tipo nel parlato e nello scritto*, 185-195.
- Voghera, M. (2017). *Dal parlato alla grammatica*. Roma: Carocci.
- Wichmann, S. (2008). The emerging field of language dynamics. *Language and Linguistics Compass*, 2(3), 442-455. <https://doi.org/10.1111/j.1749-818X.2008.00062.x>
- Wichmann, S., Rama, T., & Holman, E. W. (2011). Phonological diversity, word length, and population sizes across languages: The ASJP evidence.
- Wright, S. (1943). Isolation by distance. *Genetics*, 28(2), 114.

Sitografia

[https://www.treccani.it/enciclopedia/tipologia-linguistica_\(Enciclopedia-Italiana\)/](https://www.treccani.it/enciclopedia/tipologia-linguistica_(Enciclopedia-Italiana)/)

[https://www.treccani.it/enciclopedia/lingua-parlata_\(Enciclopedia-dell'Italiano\)/](https://www.treccani.it/enciclopedia/lingua-parlata_(Enciclopedia-dell'Italiano)/)

[https://www.treccani.it/enciclopedia/lingua-scritta_\(Enciclopedia-dell'Italiano\)/](https://www.treccani.it/enciclopedia/lingua-scritta_(Enciclopedia-dell'Italiano)/)

<https://wals.info/languoid/samples/200>

<https://www.ethnologue.com/>

<https://grambank.clld.org/>

<https://glottolog.org/>

<https://catalogue.nla.gov.au/catalog>

[https://www.treccani.it/enciclopedia/tipologia-linguistica_\(Enciclopedia-Italiana\)/](https://www.treccani.it/enciclopedia/tipologia-linguistica_(Enciclopedia-Italiana)/)

https://www.oxfordlearnersdictionaries.com/definition/english/bias_1#:~:text=%2F%C
[B%88ba%C9%AA%C9%99s%2F,not%20based%20on%20fair%20judgement](https://www.oxfordlearnersdictionaries.com/definition/english/bias_1#:~:text=%2F%C)

<https://www.unesco.org/en>

<https://www.sil.org/literacy-education>

<https://www.sil.org/language-programs-management>

Appendice

Tabella S1: Campione lingue CON tradizione scritta

#	Nome	Genus	Famiglia	Countries	Content
1	Adyghe	Circassian	Northwest Caucasian	Russia	Literature. Newspapers. Radio. TV. Dictionary. Grammar. NT
2	Ainu	Ainu	Ainu	Giappone	Newspapers. Dictionary. Grammar. Texts. NT
3	Abau	Abau	Sepik	Papua Guinea	Nuova Literature. Grammar. Texts. NT
4	Nobonob	Mabuso	Trans-New Guinea	Papua Guinea	Nuova Literature. Dictionary. NT
5	Yucuna	Western Nawiki	Maipurean	Colombia	Literature. Dictionary. Grammar. NT
6	Arabic (Egyptian)	Semitic	Afro-Asiatic	Egitto	Radio. TV. Videos. Dictionary. Grammar. NT
7	Kombio	Kombio- Arapesh	Toriccelli	Papua Guinea	Nuova Literature. Bible portions
8	Armenian Eastern	Armenian	Indo-European	Armenia	Fully developed
9	Asmat Central	Asmat- Kamrau Bay	Asmat-Kamrau Bay	Indonesia	Dictionary. Grammar. Texts. NT
10	Awa-Pit	Barbacoan	Barbacoan	Colombia	Literature. Grammar. Texts. NT
11	Aymara (Central)	Aymaran	Aymaran	Bolivia	Literature. Radio. Dictionary. Grammar. Texts. Bible
12	Bambara	Western Mande	Mande	Mali	Literature. Newspapers. Radio. TV. Videos. Dictionary. Grammar. Texts. Bible
13	Waimaha	Tucanoan	Tucanoan	Colombia	Literature. NT: 2010.
14	Basque	Basque	Basque	Spain	Fully developed
15	Gayo	Northwest Sumatra- Barrier Islands	Austronesian	Indonesia	Literature. Newspapers. Dictionary. Grammar. Texts.
16	Bawm	Kuki-Chin	Sino-Tibetan	India	Literature. Radio. Grammar. Bible
17	Beja	Beja	Afro-Asiatic	Sudan	Literature. Radio. Dictionary. Grammar. Texts. Bible portions
18	Berber (Middle Atlas)	Berber	Afro-Asiatic	Marocco	Literature. Radio. TV. Videos. Dictionary. Grammar. Texts. Bible portions

19	Brahui	Dravidian	Dravidian	Pakistan	Newspapers. Dictionary. Grammar. Texts
20	Burmese	Burmese-Lolo	Sino-Tibetan	Mynamar	Fully developed
21	Burushaski	Burushaski	Burushaski	Pakistan	Literature. Dictionary. Grammar. Texts.
22	Hopi	Northern Uto-Aztecan	Uto-Aztecan	USA	Literature. Dictionary. Grammar. NT
23	Kayapó	Je Setentrional	Macro-Ge	Brasile	Literature. Grammar. NT
24	Carib	Cariban	Cariban	Venezuela	Literature. Dictionary. Grammar. Texts. NT
25	Chamorro	Chamorro	Austronesian	Guam	Literature. Dictionary. Grammar. Texts. NT
26	Chinantec, Comaltepec	Chinantecan	Oto-Manguean	Messico	Literature. Radio. Videos. Texts. NT
27	Chukchi	Northern Chukotko-Kamchatkan	Chukotko-Kamchatkan	Russia	Dictionary. Grammar. Texts. Bible portions
28	Nahuatl, Eastern Huasteca	Southern Uto-Aztecan	Uto-Aztecan	Messico	Radio. Dictionary. Grammar. Bible
29	Cree (Plains)	Algonquian	Algic	Canada	Dictionary. Grammar. Texts. Bible
30	Daga	Dagan	Trans-New Guinea	Papua Nuova Guinea	Literature. Grammar. NT
31	Yali, Ninia	Ngalik	Trans-New Guinea	Indonesia	Literature. Bible
32	Jola-Kasa	Jola	Niger-Congo	Senegal	Literature. Radio. Dictionary. Grammar. NT
33	Iaai	Oceanic	Austronesian	New Caledonia	Literature. Dictionary. Grammar. Bible
34	English	Germanic	Indo-European	Regno Unito	Fully developed.
35	Emberá, Northern	Choco	Choco	Colombia	Literature. Radio. Dictionary. Grammar. NT
36	Evenki	Tungusic	Altaic	Cina	Radio. TV. Dictionary. Grammar. Texts. Bible portions
37	Ewe	Gbe	Niger-Congo	Ghana	Literature. Newspapers. Radio. TV. Dictionary. Grammar. Texts. Bible
38	Fijian	Oceanic	Austronesian	Fiji	Newspapers. Radio. Dictionary. Grammar. Texts. Bible
39	Finnish	Finnic	Uralic	Finlandia	Fully developed.
40	French	Romance	Indo-European	Francia	Fully developed.
41	Fur	Fur	Fur	Sudan	Literature. Radio. Dictionary. Grammar. Texts.
42	Garo	Brahmaputran	Sino-Tibetan	India	Literature. Radio. Videos. Dictionary. Grammar. Bible
43	Georgian	Kartvelian	Kartvelian	Georgia	Fully developed.
44	German	Germanic	Indo-European	Germania	Fully developed.

45	Greek (Modern)	Greek	Indo-European	Grecia	Fully developed.
46	Greenlandic (West)	Eskimo	Eskimo-Aleut	Greenland	Literature. Radio. Dictionary. Grammar. Texts. Bible
47	Guaraní Mbyá	Tupi-Guarani	Tupian	Brasile	Literature. Dictionary. Grammar. Texts. Bible
48	Tainae	Nuclear Angan	Trans-New Guinea	Papua Guinea	Nuova Literature.
49	Hausa	West Chadic	Afro-Asiatic	Nigeria	Literature. Radio. TV. Dictionary. Grammar. Bible
50	Hebrew (Modern)	Semitic	Afro-Asiatic	Israele	Fully developed
51	Hindi	Indic	Indo-European	India	Fully developed.
52	Yukpa	Cariban	Cariban	Colombia	Literature. Dictionary. Texts.
53	Hmong Njua	Hmongic	Hmong-Mien	Cina	Literature. Dictionary. Grammar. Bible
54	Hungarian	Ugric	Uralic	Ungheria	Fully developed.
55	Igbo	Igboid	Niger-Congo	Nigeria	Literature. Radio. TV. Dictionary. Grammar. Bible
56	Kogi	Arhuacic	Chibchan	Colombia	Literature. Texts. NT
57	Waris	Border	Border	Papua Guinea	Nuova Literature. Dictionary. Texts. NT
58	Indonesian	Malayo-Sumbawan	Austronesian	Indonesia	Fully developed
59	Chechen	Nakh	Nakh-Daghestanian	Russia	Newspapers. Radio. Dictionary. Grammar. Bible
60	Alagwa	Southern Cushitic	Afro-Asiatic	Tanzania	Literature. Grammar. Bible portions
61	Irish	Celtic	Indo-European	Irlanda	Radio. Dictionary. Grammar. Texts. Bible
62	Q'eqchi'	Mayan	Mayan	Guatemala	Literature. Radio. TV. Videos. Dictionary. Grammar. Texts. Bible
63	Japanese	Japanese	Japanese	Giappone	Fully developed.
64	Kannada	Dravidian	Dravidian	India	Fully developed.
65	Kanuri	Western Saharan	Saharan	Nigeria	Radio. TV. Dictionary. Grammar. Texts. NT
66	Zuni	Zuni	Zuni	USA	Radio. Videos. Dictionary. Grammar. Bible
67	Karen, Geba	Karen	Sino-Tibetan	Myanmar	Literature. Radio. Grammar. Bible portions
68	Dangaléat	East Chadic	Afro-Asiatic	Chad	Literature. Dictionary. Grammar. Texts. NT
69	Khalkha	Mongolic	Altaic	Mongolia	Literature. Newspapers. Periodicals. Radio. TV. Videos. Dictionary. Grammar. Texts. Bible
70	Khasi	Khasian	Austro-Asiatic	India	Literature. Radio. TV. Dictionary. Grammar. Bible

71	Khmer	Khmer	Austro-Asiatic	Cambogia	Fully developed.
72	Khmu'	Khmuic	Austro-Asiatic	Laos	Radio. TV. Dictionary. Grammar. Bible portions
73	Misima-Panaeati	Oceanic	Austronesian	Papua Nuova Guinea	Dictionary. NT. Literacy rate in L1: 75%–80%.
74	Carolinian	Oceanic	Austronesian	Northern Mariana Islands	Literature. Newspapers. Videos. Dictionary. Texts.
75	Kobon	Kalam-Kobon	Trans-New Guinea	Papua Nuova Guinea	Literature. Grammar. NT
76	Kongo	Bantu	Niger-Congo	Rep. Democratica Congo	Radio. TV. Dictionary. Grammar. Bible
77	Korean	Korean	Korean	Corea del sud	Fully developed.
78	Koromfe	Koromfe	Niger-Congo	Burkina Faso	Literature. Dictionary. Grammar. Texts. NT
79	Koyraboro Senni	Songhay	Songhay	Mali	Literature. Dictionary. Grammar. NT
80	Katcha-Kadugli-Miri	Kadugli	Kadu	Sudan	Literature. Grammar. NT
81	Kunama	Kunama	Kunama	Eritrea	Literature. Radio. Videos. Dictionary. Grammar. Texts. Bible
82	Kutenai	Kutenai	Kutenai	Canada	Dictionary. Grammar. Texts
83	Ladakhi	Bodic	Sino-Tibetan	India	Literature. Dictionary. Grammar. Bible
84	Lak	Lak	Nakh-Daghestanian	Russia	Literature. Newspapers. Periodicals. Radio. Videos. Dictionary. Grammar. Bible portions
85	Lango	Western Nilotic	Eastern Sudanic	Uganda	Literature. Newspapers. Radio. Dictionary. Grammar. Texts. Bible
86	Latvian	Baltic	Indo-European	Latvia	Fully developed.
87	Hruso	Himalayish	Sino-Tibetan	India	Literature.
88	Lezgian	Lezxic	Nakh-Daghestanian	Russia	Literature. Newspapers. Radio. Dictionary. Grammar. Texts. NT
89	Luvale	Bantu	Niger-Congo	Zambia	Newspapers. Radio. TV. Grammar. Bible
90	Malagasy	Barito	Austronesian	Madagascar	Fully developed.
91	Mandarin	Chinese	Sino-Tibetan	Cina	Fully developed.
92	Tahitian	Oceanic	Austronesian	Polinesia francese	Literature. Periodicals. Radio. TV. Videos. Dictionary. Grammar. Texts. Bible
93	Meithei	Manipuri	Sino-Tibetan	India	Literature. Radio. Videos. Dictionary. Grammar. Bible
94	Mixtec (Chalcatongo)	Mixtec	Oto-Manguenan	Messico	Literature. Dictionary. Grammar. NT

95	Mundari	Munda	Austro-Asiatic	India	Literature. Periodicals. Radio. Dictionary. Grammar. Texts. Bible
96	Murle	South Surmic	Eastern Sudanic	Sudan del sud	Literature. Dictionary. Grammar. NT
97	Nahuatl Eastern Durango	Southern Uto-Aztecan	Uto-Aztecan	Messico	Literature. Radio. Videos. Texts. Bible portions
98	Nama	Khoe-Kwadi	Khoe-Kwadi	Namibia	Radio. Dictionary. Grammar. Bible
99	Ndyuka	Creoles and Pidgins	other	Suriname	Literature. Radio. Dictionary. Grammar. Texts. NT
100	Nenets (Forest)	Samoyedic	Uralic	Russia	Literature. Radio. TV. Dictionary. Grammar. Texts. Bible portions
101	Lendu	Lendu	Central Sudanic	Repubblica Decocratica del Congo	Literature. Dictionary. Texts. Bible
102	Arrernte, Eastern	Arandic	Pama-Nyungan	Australia	Radio. TV. Videos. Dictionary. Grammar. Texts. NT
103	Nkore-Kiga	Bantu	Niger-Congo	Uganda	Literature. Newspapers. Radio. Dictionary. Grammar. Bible
104	Nobiin	Nubian	Eastern Sudanic	Sudan	Literature. Dictionary. Grammar. Bible portions
105	Wyandot	Northern Iroquoian	Iroquoian	USA	Dictionary. Grammar. Audio. Video. Narratives
106	Oromo, West Central	Lowland East Cushitic	Afro-Asiatic	Etiopia	Literature. Newspapers. Radio. TV. Dictionary. Grammar. Bible
107	Otomi, Eastern Highland	Otomian	Oto-Manguean	Messico	Literature. Dictionary. Grammar. NT
108	Apma	Oceanic	Austronesian	Venuatu	Literature. Dictionary. Grammar. Bible portions
109	Meskwaki	Algonquian	Algie	USA	Literature. Dictionary. Grammar. Bible portions
110	Kulina	Arauan	Arauan	Brasile	Literature. Dictionary. Grammar. Texts. NT
111	Persian	Iranian	Indo-European	Iran	Fully developed.
112	Quichua, Chimborazo Highland	Quechuan	Quechuan	Ecuador	Literature. Dictionary. Grammar. Texts. Bible
113	Rapanui	Oceanic	Austronesian	Cile	Literature. Radio. Dictionary. Grammar. NT
114	Russian	Slavic	Indo-European	Russia	Fully developed.
115	Semai	Aslian	Austro-Asiatic	Malesia	Literature. Dictionary. Bible portions

116	Shipibo-Konibo	Panoan	Pano-Tacanan	Peru		Literature. Radio. Dictionary. Grammar. NT
117	Gwich'in	Athapaskan	Na-Dene	Canada		Literature. Dictionary. Bible
118	Spanish	Romance	Indo-European	Spagna		Fully developed.
119	Binandere	Binanderean	Trans-New Guinea	Papua Guinea	Nuova	Literature. Dictionary. Bible portions
120	Supyire	Senufo	Niger-Congo	Mali		Literature. Radio. Videos. Grammar. NT
121	Swahili	Bantu	Niger-Congo	Tanzania		Fully developed.
122	Tagalog	Greater Central Philippine	Austronesian	Filippine		Fully developed.
123	Thai	Kam-Tai	Tai-Kadai	Tailandia		Fully developed.
124	Cia-Cia	Celebic	Austronesian	Indonesia		Literature. Videos.
125	Tunica	Tunica	Tunica	USA		Literature. Dictionary. Grammar. Texts.
126	Turkish	Turkic	Altaic	Turchia		Fully developed.
127	Aché	Tupi-Guarani	Tupian	Paraguay		Literature. Radio. Dictionary. Texts. NT
128	Girawa	Kokon	Trans-New Guinea	Papua Guinea	Nuova	Literature. Dictionary. Grammar. NT
129	Vietnamese	Vietic	Austro-Asiatic	Vietnam		Fully developed.
130	Wichí	Matacoan	Matacoan	Argentina		Radio. Texts. Bible
131	Bosmun	Ramu	Ramu-Lower Sepik	Papua Guinea	Nuova	Literature.
132	Yoruba	Defoid	Niger-Congo	Nigeria		Newspapers. Radio. TV. Dictionary. Grammar. Bible
133	Tunica	Tunica	Tunica	USA		Literature. Dictionary. Grammar. Texts.
134	Yup'ik (Central)	Eskimo	Eskimo-Aleut	USA		Literature. Dictionary. Grammar. Bible
135	Yurok	Yurok	Algic	USA		Dictionary. Grammar. Texts
136	Popoluca, Highland	Zoquean	Mixe-Zoque	Messico		Literature. Radio. Dictionary. NT
137	Zulu	Bantu	Niger-Congo	Sud Africa		Newspapers. Radio. Dictionary. Grammar. Bible
138	Aringa	Moru-Mari	Central Sudanic	Uganda		Literature. Radio. Texts. NT
139	Waorani	Waorani	Waorani	Ecuador		Literature. Grammar. Texts. NT
140	Avar	Avar-Andic-Tsezic	Nakh-Daghestanian	Russia		Literature. Newspapers. Radio. TV. Dictionary. Grammar. Texts. NT
141	Masalit	Maban	Nilo-Saharan	Sudan		Literature. Radio. Grammar. Texts.
142	Mapudungun	Araucanian	Araucanian	Cile		Radio. Dictionary. Grammar. Texts. NT
143	Cocopa	Yuman	Hokan	Messico		Dictionary. Grammar. Texts. Bible portions
144	Kashaya	Pomoan	Hokan	USA		Dictionary. Grammar. Texts.

145	Tlingit	Tlingit	Na-Dene	USA	Radio. Grammar. portion	Dictionary. Bible
146	Tsimshian (Coast)	Tsimshianic	Tsimshianic	Canada	Dictionary. Bible portions	Grammar.
147	Tarahumara, Western	Tarahumara	Uto-Aztecan	Messico	Literature. NT	

Tabella S2: Campione lingue SENZA tradizione scritta

#	Nome	Genus	Famiglia	Countries	Content
1	Mocoví	Guaicuru del sur	Guaicuruan	Argentina	Grammar. NT
2	Keres Eastern	Keresan	Keresan	Stati Uniti d'America	No info
3	Kaningra	Sepik Hill	Sepik	Papua Nuova Guinea	No info
4	Gumalu	Mabuso	Trans-New Guinea	Papua Nuova Guinea	No info
5	Iñapari	Purus	Maipurean	Perù	Dictionary.
6	Arabic, Spoken	Sa'idi Semitic	Afro-Asiatic	Egitto	Grammar.
7	Reyesano	Tacanan	Pano-Tacanan	Bolivia	unwritten
8	Abu'	Kombio- Arapesh	Torricelli	Papua Nuova Guinea	Grammar. (limited usage of writing)
9	Asmat North	Asmat- Kamrau Bay	Trans-New Guinea	Indonesia	No info
10	Morom	Bongo- Bagirmi	Central Sudanic, Nilo-Saharan	Ciad	unwritten
11	Maninkakan, Kita	Manding-West	Niger-Congo (Mande)	Mali	Grammar.
12	Macaguaje	Tucanoan	Tucanoan	Colombia	No info
13	Batak Kluet	Alas- Northwest Sumatra- Barrier Islands	Austronesian	Indonesia	Dictionary.
14	Purum	Kuki-Chin	Sino-Tibetana	India	No info
15	Dahalik	Ethiopian	Afro-Asiatica	Eritrea	unwritten
16	Senhaja Berber	Berber	Afro-Asiatica	Marocco	Dictionary.
17	Bribri	Western Isthmic Chibchan	Chibchan	Costa Rica	Dictionary.Grammar.NT
18	Tübatulabal	Northern Uto- Aztecan	Uto-Atzecan	Stati Uniti d'America	Dictionary. Grammar.
19	Gavião, Pará	Je Setentrional	Macro-Ge	Brasile	Dictionary. Grammar.
20	Chaima	Cumanà	Cariban	Venezuela	Dictionary.

21	Cayuvava	Cayuvava	Cayuvava	Bolivia		Dictionary.Texts. Unwritten
22	Chinantec, Chiltepec	Chinantecan	Otomanguean	Messico		unwritten
23	Alutor	Koryak- Alyutor	Chukotko- Kamchatkan	Russia		Dictionary.
24	Timbisha	Northern Uto- Aztecan	Uto-Aztecan	Stati Uniti d'America		Dictionary.Grammar.
25	Coos	Coosan	Oregon Coast	Stati Uniti d'America		Grammar.Texts.
26	Ginuman	Dagan	Trans-New Guinea (Dagan)	Papua Nuova Guinea		unwritten
27	Dani	Dani	Trans-New Guinea (Dani)	Indonesia		Grammar.NT
28	Kerak	Jola	Niger-Congo	Senegal		unwritten
29	Haeke	Oceanic	Austronesian	Nuova Caledonia		No info
30	Dao	Paniai Lakes	Trans-new Guinea	Indonesia		NT
31	Embera Baudò	Choco	Choco	Colombia		No info
32	Oroquen	Tungusic	Altaic	Cina		radio
33	Gbe Western Xwla	Gbe	Niger-Congo	Benin		unwritten
34	Gone Dau	Oceanic	Austronesian	Figi		No info
35	Ingrian	Finnic	Uralic	Russia		Dictionary.Grammar.
36	Amdang	Fur	Nilo-Saharan	Ciad		unwritten
37	Tiwa	Brahmaputran	Sino-Tibetana	India		Dictionary.Grammar.
38	Bunuba	Bunuban	Bunuban	Australia		Grammar.
39	Grebo	Kru	Niger-Congo	Liberia		Dictionary.Grammar. unwritten
40	Cappadocian Greek	Greek	Indoeuropean	Grecia		Dictionary.Grammar.
41	North Alaskan Inupiatun	Eskimo	Eskimo-Aleut	Stati Uniti d'America		Dictionary.Grammar. NT
42	Pai Tavytera	Guarani	Tupian	Paraguay		No info
43	Haida southern	Haida	Haida	Canada		Dictionary.Grammar.songs texts
44	Kamasa	Nuclear- Angan	Trans-New Guinea	Papua Nuova Guinea		No info

45	Gwandara	West-Chadic	Afro-Asiatic	Nigeria	Texts.
46	Sansi	Indic	Indoeuropean	India	unwritten
47	Kaxuiàna	Cariban	Cariban	Brasile	unwritten
48	Nonuya	Witoto	Witotan	Colombia	unwritten
49	Hunzib	Avar-Andic- Tsezic	Nakh- Daghestanian	Georgia	Dictionary.Grammar. Texts. unwritten
50	Ukwuani- Aboh-Ndoni	Igboid	Niger-Congo	Nigeria	Grammar.
51	Malayo	Guamaca- Atanque	Chibchan	Colombia	No info
52	Imonda	Border	Border	Papua Nuova Guinea	Grammar.
53	Col	Malay	Austronesian	Indonesia	unwritten
54	Akhvakh	Andic	Nakh- Daghestanian	Russia	Dictionary. Grammar. unwritten
55	Gorowa	Southern Cushitic	Afro-Asiatic	Tanzania	Dictionary. Texts. Unwritten
56	Itza'	Mayan	Mayan	Guatemala	Dictionary. Grammar. Texts.
57	Amami- Oshima, Northern	Ryukyuan	Japonic	Giappone	No info
58	Northwestern !Kung	Ju-Kung	Kxa	Angola	radio, unwritten
59	Urali	Dravidian	Dravidian	India	unwritten
60	Kanembu	Western Saharan	Saharan	Ciad	Dictionary. Grammar. Texts. uso marginale della scrittura
61	Chitimacha	Chitimacha	Chitimacha	Stati Uniti d'America	Grammar. Texts.
62	Karen geko	Karen	Sino-Tibetana	Birmania	Dictionary.
63	Kayardild	Tangkic	Tangkic	Australia	Dictionary. Grammar. Texts.
64	Kwang	East Chadic	Afro-asiatic	Chad	No info
65	Ket	Yeniseian	Yeniseian	Russia	Dictionary. Grammar.
66	Lembena	Enga	Trans-New- Guinea	Papua Nuova Guinea	No info
67	Khamnigan Mongol	Mongolic	Altaic	China, Mongolia, Russia	No info

68	Lyngngam	Khasian	Austro-Asiatic	India	No info
69	Kachok	Bahnaric	Austro-Asiatic	Cambogia	unwritten
70	Khuen	khmuic	Austro-Asiatic	Laos	No info
71	Budibud	Oceanic	Austronesian	Papua Nuova Guinea	No info
72	Jemez	Kiowa-Tanoan	Kiowa-Tanoan	USA	No info
73	Namonuito	Oceanic-Micronesian	Austronesian	Micronesia	No info
74	Koasati	Muskogean	Muskogean	USA	Dictionary. Grammar. Bible. (uso minimo della scrittura)
75	Wagama	Kalam-kobon	Trans-New-Guinea	Papua Nuova Guinea	No info
76	Kaamba	Bantu	Niger-congo	Congo	unwritten
77	Bwamu, Cwi	Gur	Niger-congo	Birkina Faso	No info
78	Tondi Songway Kiini	Songhay	Nilo-Saharan	Mali	Dictionary. Grammar. unwritten
79	Kanga	Kadugli	Kadu	Sudan	No info
80	Changthang	Bodic	Sino-Tibetan	India	No info
81	Kubachi	Dargi	Nakh-Daghestanian	Russia	Grammar.
82	Quapaw	Mississippi Valley Siouan	Siouan	USA	Dictionary.
83	Dongotono	Nilotic	Eastern Sudanic	South Sudan	unwritten
84	Lavukaleve	Lavukaleve	Solomons East Papuan	Solomon Islands	Grammar.
85	Budukh	Lezgif	Nakh-Daghestanian	Azerbaijan	Dictionary. Grammar. unwritten
86	Mbwela	Bantu	Niger-Congo	Angola	No info
87	Marfa	Mabian	Niger-Congo	Chad	No info
88	Makah	Southern Wakashan	Wakashan	USA	Grammar.
89	Bushi	Barito	Autronesian	Mayotte	No info
90	Jinyu Chinese	Chinese	Sino-Tibetan	Cina	Grammar.

91	Mangarrayi	Mangarrayi- Maran	Mangarrayi- Maran	Australia	Dictionary. Texts.	Grammar.
92	Niufo'ou	Oceanic	Austronesian	Tonga	Grammar.	
93	Huilliche	Araucanian	Araucanian	Cile	Dictionary. Texts. unwritten	Grammar.
94	Maranunggu	Wagaydy	Western Daly	Australia	Texts.	
95	Maricopa	Yuman	Hokan	USA	unwritten	
96	Marind	Marind-Yaqay	Trans-New Guinea	Indonesia	Dictionary. Grammar.	
97	Kurrama	Western Pama- Nyungan	Pama- Nyungan	Australia	Dictionary. unwritten	Texts.
98	Amurdak	Amaragic	Iwaidjan	Australia	Dictionary. Texts.	
99	Karon Dori	Maybrat	Maybrat	Indonesia	No info	
100	Miwok (Southern Sierra)	Miwok	Penutian	USA	Grammar.	
101	Mixtec, Yutanduchi	Mixtec	Oto- Manguean	Messico	No info	
102	Asuri	Munda	Austro- Asiatic	India	No info	
103	Kacipo-Bale Suri	South Surmic	Eastern Sudanic	South Sudan	Dictionary.	
104	Nahuatl, Tabasco	Nahuatl	Uto-Atzecan	Messico	unwritten	
105	Xiri	Khoe-Kwadi	Khoe-Kwadi	Sudafrica	No info	
106	Apache, Mescalero- Chiricahua	Apachean	Na-Dene	USA	Dictionary. unwritten	Texts.
107	Kwinti	Creoles and Pidgins	other	Suriname	unwritten	
108	Nganasan	Samoyedic	Uralic	Russia	Dictionary. Grammar.	
109	Tenino	Sahaptian	Penutian	USA	No info	
110	Ngiti	Lendu	Central Sudanic	Repubblica democratica del Congo	Grammar. NT	
111	Gamilaraay	Southeastern Pama- Nyungan	Pama- Nyungan	Australia	Dictionary. Grammar.	
112	Nivkh	Nivkh	Nivkh	Russia	Dictionary. Grammar.	
113	Kenye	Bantu	Niger-Congo	Uganda	radio, unwritten	

114	Nubian (Dongolese)	Nubian	Eastern Sudanic	Sudan	Dictionary. Grammar.	
115	Ngandi	Enindhilyagwa	Gunwinyguan	Australia	Dictionary. Texts.	Grammar.
116	Onondaga	Northern Iroquoian	Iroquoian	Canada	Dictionary.	
117	Oromo (Harar)	Lowland East Cushitic	Afro-Asiatic	Etiopia	Dictionary. Texts.	Grammar.
118	Otomi Ixtenco	Otomian	Oto- Manguean	Messico	unwritten	
119	Lehali	Oceanic	Austronesian	Vanuatu	No info	
120	Suruahá	Arauan	Arauan	Brasile	No info	
121	Dehwari	Iranian	Indo- European	Pakistan	unwritten	
122	Pirahã	Mura	Mura	Brasile	Grammar. Bible.	
123	Yankunytjatjara	Western Pama- Nyungan	Pama- Nyungan	Australia	Dictionary. Texts. unwritten	Grammar.
124	Pomo (Southeastern)	Pomoan	Hokan	USA	Grammar.	
125	Qawasqar	Alacalufan	Alacalufan	Cile	Dictionary. Texts. unwritten	Grammar.
126	Quichua, Calderón Highland	Quechuan	Quechuan	Ecuador	No info	
127	Rama	Rama	Chibchan	Nicaragua	Dictionary. unwritten	Grammar.
128	Sango Riverain	Ubangi	Niger-Congo	Repubblica Centrafricana	unwritten	
129	Sanuma	Yanomam	Yanomam	Venezuela	Grammar. NT	
130	Tehuelche	Chonan Proper	Chonan	Argentina	Dictionary. unwritten	Grammar.
131	Temoq	Aslian	Austro- Asiatic	Malesia	unwritten	
132	Nafri	Sentanic	Sentanic	Indonesia	No info	
133	Katukína, Panoan	Panoan	Pano-Tacanan	Brasile	Grammar.	
134	Slave	Athapaskan	Na-Dene	Canada	Dictionary. Grammar.	
135	Comox	Central Salish	Salishan	Canada	Grammar.	
136	Gaina	Binanderean	Trans-New Guinea	Papua Nuova Guinea	No info	

137	Sénoufo, Nanerigé	Senufo	Niger-Congo	Burkina Faso	No info	
138	Doe	Narrow Bantu	Niger-Congo	Tanzania	unwritten	
139	Taba	South Halmahera - West New Guinea	Austronesian	Indonesia	Grammar. unwritten	
140	Mandaya	Greater Central Philippine	Austronesian	Filippine	Dictionary. unwritten	
141	Nyaw	Kam-Tai	Tai-Kadai	Tailandia	No info	
142	Trumai	Trumai	Trumai	Brasile	No info	
143	Gitksan	Nass-Gitksan	Tsimshianic	Canada	Dictionary. Bible.	Grammar.
144	Tukang Besi	Celebic	Austronesian	Indonesia	Dictionary. unwritten	Grammar.
145	Balkan Gagauz Turkish	Turkic	Altaic	Turchia	Texts.	
146	Ungarinjin	Worrorran	Worrorran	Australia	Dictionary. Texts.	Grammar.
147	Guajà	Maweti- Guarani	Tupian	Brasile	Grammar.	
148	Yaben	North Adelbert	Trans-New Guinea	Papua Nuova Guinea	No info	
149	Chut	Vietic	Austro- Asiatic	Vietnam	unwritten	
150	Wambaya	Wambayan	Mirndi	Australia	Dictionary. Texts.	Grammar.
151	Yangman	Yangmanic	Yangmanic	Australia	No info	
152	Oro Win	Chapacura- Wanham	Chapacura- Wanham	Brasile	unwritten	
153	Bermejo Wichí	Matacoan	Matacoan	Argentina	Dictionary. Grammar.	
154	Wichita	Northern Caddoan	Caddoan	USA	Grammar.	
155	Yidyni	Northern Pama- Nyungan	Pama- Nyungan	Australia	Dictionary. Texts. unwritten	Grammar.
156	Yimas	Panoan	Ramu-Lower Sepik	Papua Nuova Guinea	Grammar.	
157	Ulukwumi	Defoid	Niger-Congo	Nigeria	No info	

158	Yuchi	Yuchi	Yuchi	USA	Grammar. unwritten
159	Yukaghir (Kolyma)	Yukaghir	Yukaghir	Russia	Dictionary. Grammar.
160	Yupik lawrence island	st Eskimo	Eskimo-Aleut	USA	Dictionary. Grammar. Texts. NT
161	Zoque, Chimalapa	Mixe-Zoque	Mixe-Zoque	Messico	Grammar.
162	Birwa	Bantu	Niger-Congo	Botswana	unwritten
163	Jaquaru	Aymaran	Tupe	Perù	Dictionary. Grammar. Texts.
164	Maninka, Konyanka	Manding	Mande	Guinea	unwritten
165	Kumarbhag Paharia	Dravidian	Dravidian	India	Dictionary. Grammar. unwritten
166	Danu	Burmese-Lolo	Sino-Tibetan	Myanmar	unwritten
167	Menominee	Algonquian	Algic	USA	Dictionary. Grammar.
168	Miao, Western Mashan	Hmongic	Hmong-Mien	China	unwritten
169	Lish	Central Tibeto- Burman	Sino-Tibetan	India	unwritten
170	Narragansett	Algonquian	Algic	USA	unwritten
171	Una	Mek	Trans-New Guinea	Indonesia	Dictionary. NT
172	Yaqui	Cahita	Uto-Aztecan	Messico	Dictionary. Grammar. NT
173	Saxon, East Frisian Low	Germanic	Indoeuropean	Germania	Dictionary. Grammar.
174	Mirandese	Romantic	Indoeuropean	Portogallo	Bible portions.
175	Saami, Ume	Sami	Uralic	Sweden	Bible.
176	Acoma	Keresan	Keresan	USA	Grammar. Texts. Bible.

Tabella S3: Campione WALS

#	Nome	Genus	Famiglia	Countries	Content
1	Abipón	Abipon	Guaicuruan	Argentina	Grammar
2	Abkhaz	Northwest Caucasian	Northwest Caucasian	Georgia	Dictionary, grammar, NT
3	Acoma	Keresan	Keresan	USA	Dictionary, grammar, NT
4	Ainu	Ainu	Ainu	Giappone	Newspapers. Dictionary. Grammar. Texts. NT
5	Alamblak	Sepik Hill	Sepik	Papua Guinea	Nuova Grammar. NT
6	Amele	Mabuso	Trans-New Guinea	Papua Guinea	Nuova Grammar. NT
7	Apurinã	Purus	Arawakan	Brasile	Grammar. Texts. NT
8	Arabic (Egyptian)	Semitic	Afro-Asiatic	Egitto	Radio. TV. Videos. Dictionary. Grammar. NT
9	Araona	Tacanan	Pano-Tacanan	Bolivia	Dictionary. Grammar. Bible
10	Arapesh (Mountain)	Kombio-Arapesh	Torricelli	Papua Guinea	Nuova Grammar. NT
11	Armenian (Eastern)	Armenian	Indo-European	Armenia	Fully developed
12	Asmat	Asmat-Kamrau Bay	Asmat-Kamrau Bay	Indonesia	Dictionary. Grammar. Texts. NT
13	Awa Pit	Barbacoan	Barbacoan	Colombia	Literature. Grammar. Texts. NT
14	Aymara (Central)	Aymaran	Aymaran	Bolivia	Literature. Radio. Dictionary. Grammar. Texts. Bible
15	Bagirmi	Bongo-Bagirmi	Central Sudanic	Chad	Dictionary. Grammar. Bible portions
16	Bambara	Western Mande	Mande	Mali	Literature. Newspapers. Radio. TV. Videos. Dictionary. Grammar. Texts. Bible
17	Barasano	Tucanoan	Tucanoan	Colombia	Dictionary. Grammar. Texts. NT
18	Basque	Basque	Basque	Spain	Fully developed
19	Batak (Karo)	Northwest Sumatra-Barrier Islands	Austronesian	Indonesia	Dictionary. Grammar. Bible
20	Bawm	Kuki-Chin	Sino-Tibetan	India	Literature. Radio. Grammar. Bible

21	Beja		Beja	Afro-Asiatic	Sudan	Literature. Radio. Dictionary. Grammar. Texts. Bible portions
22	Berber (Middle Atlas)		Berber	Afro-Asiatic	Marocco	Literature. Radio. TV. Videos. Dictionary. Grammar. Texts. Bible portions
23	Brahui		Dravidian	Dravidian	Pakistan	Newspapers. Dictionary. Grammar. Texts
24	Bribri		Western Isthmic Chibchan	Chibchan	Costa Rica	Dictionary, grammar, NT
25	Burmese		Burmese-Lolo	Sino-Tibetan	Myanmar	Fully developed
26	Burushaski		Burushaski	Burushaski	Pakistan	Literature. Dictionary. Grammar. Texts.
27	Cahuilla		Northern Uto-Aztecan	Uto-Aztecan	USA	Dictionary. Grammar. Texts
28	Canela		Je Setentrional	Macro-Ge	Brasil	Grammar. NT
29	Carib		Cariban	Cariban	Venezuela	Literature. Dictionary. Grammar. Texts. NT
30	Cayuvava		Cayuvava	Cayuvava	Bolivia	Dictionary, texts, unwritten
31	Chamorro		Chamorro	Austronesian	Guam	Literature. Dictionary. Grammar. Texts. NT
32	Chinantec (Lealao)		Chinantecan	Oto-Manguéan	Mexico	Dictionary. Grammar. Texts. NT
33	Chukchi		Northern Chukotko-Kamchatkan	Chukotko-Kamchatkan	Russia	Dictionary. Grammar. Texts. Bible portions
34	Comanche		Northern Uto-Aztecan	Uto-Aztecan	USA	Dictionary. Grammar. Texts. Bible portions
35	Coos (Hanis)		Coosan	Oregon Coast	Stati Uniti d'America	Grammar, texts
36	Cree (Plains)		Algonquian	Algic	Canada	Dictionary. Grammar. Texts. Bible
37	Daga		Dagan	Trans-New Guinea	Papua Guinea Nuova	Literature. Grammar. NT
38	Dani (Lower Grand Valley)		Dani	Trans-New Guinea	Indonesia	Grammar, NT
39	Diola-Fogny		Jola	Niger-Congo	Senegal	Radio. TV. Dictionary. Grammar. NT
40	Drehu		Oceanic	Austronesian	New Caledonia	Dictionary. Grammar. Texts. Bible
41	Ekari		Paniai Lakes	Trans-New Guinea	Indonesia	Dictionary. Texts. NT
42	English		Germanic	Indo-European	Regno Unito	Fully developed.
43	Epena Pedee		Choco	Choco	Colombia	Dictionary. Grammar. NT

44	Evenki	Tungusic	Altaic	Cina	Radio. TV. Dictionary. Grammar. Texts. Bible portions
45	Ewe	Gbe	Niger-Congo	Ghana	Literature. Newspapers. Radio. TV. Dictionary. Grammar. Texts. Bible
46	Fijian	Oceanic	Austronesian	Fiji	Newspapers. Radio. Dictionary. Grammar. Texts. Bible
47	Finnish	Finnic	Uralic	Finlandia	Fully developed.
48	French	Romance	Indo-European	Francia	Fully developed.
49	Fur	Fur	Fur	Sudan	Literature. Radio. Dictionary. Grammar. Texts.
50	Garo	Brahmaputran	Sino-Tibetan	India	Literature. Radio. Videos. Dictionary. Grammar. Bible
51	Georgian	Kartvelian	Kartvelian	Georgia	Fully developed.
52	German	Germanic	Indo-European	Germania	Fully developed.
53	Gooniyandi	Bunuban	Bunuban	Australia	Dictionary. Grammar. Texts
54	Grebo	Kru	Niger-Congo	Liberia	Dictionary, Grammar, unwritten
55	Greek (Modern)	Greek	Indo-European	Grecia	Fully developed.
56	Greenlandic (West)	Eskimo	Eskimo-Aleut	Greenland	Literature. Radio. Dictionary. Grammar. Texts. Bible
57	Guaraní	Maweti-Guarani	Tupian	Brasile	Literature. Dictionary. Grammar. Texts. Bible
58	Haida	Haida	Haida	Canada	Radio. Dictionary. Grammar. Texts. Bible
59	Hamtai	Nuclear Angan	Trans-New Guinea	Papua Guinea Nuova	Grammar. NT
60	Hausa	West Chadic	Afro-Asiatic	Nigeria	Literature. Radio. TV. Dictionary. Grammar. Bible
61	Hebrew (Modern)	Semitic	Afro-Asiatic	Israele	Fully developed
62	Hindi	Indic	Indo-European	India	Fully developed.
63	Hixkaryana	Cariban	Cariban	Brasile	Grammar. Texts. NT
64	Hmong Njua	Hmongic	Hmong-Mien	Cina	Literature. Dictionary. Grammar. Bible
65	Huitoto (Minica)	Witoto	Witotoan	Colombia	Dictionary. Grammar. NT
66	Hungarian	Ugric	Uralic	Ungheria	Fully developed.

67	Hunzib	Avar-Andic-Tsezic	Nakh-Daghestanian	Georgia	Dictionary, grammar, texts, unwritten
68	Igbo	Igboid	Niger-Congo	Nigeria	Literature. Radio. TV. Dictionary. Grammar. Bible
69	Ika	Arhuacic	Chibchan	Colombia	Grammar. Texts. Bible
70	Imonda	Border	Border	Papua Guinea Nuova	Grammar
71	Indonesian	Malayo-Sumbawan	Austronesian	Indonesia	Fully developed
72	Ingush	Nakh	Nakh-Daghestanian	Russia	Radio. Dictionary. Grammar. Texts. Bible
73	Iraqw	Southern Cushitic	Afro-Asiatic	Tanzania	Dictionary. Grammar. Texts. Bible
74	Irish	Celtic	Indo-European	Irlanda	Radio. Dictionary. Grammar. Texts. Bible
75	Jakaltek	Mayan	Mayan	Guatemala	Dictionary. Grammar. Texts. Bible
76	Japanese	Japanese	Japanese	Giappone	Fully developed.
77	Ju 'hoan	Ju-Kung	Kxa	Namibia	Dictionary. Grammar. NT
78	Kannada	Dravidian	Dravidian	India	Fully developed.
79	Kanuri	Western Saharan	Saharan	Nigeria	Radio. TV. Dictionary. Grammar. Texts. NT
80	Karok	Karok	Hokan	USA	Grammar. Text.
81	Kayah (Eastern)	Li Karen	Sino-Tibetan	Myanmar	Grammar. Texts
82	Kayardild	Tangkic	Tangkic	Australia	Dictionary, grammar, texts
83	Kera	East Chadic	Afro-Asiatic	Chad	Dictionary. Grammar. Texts. NT
84	Ket	Yeniseian	Yeniseian	Russia	Dictionary. Grammar
85	Kewa	Enga_Kewa-Huli	Trans-New Guinea	Papua Guinea Nuova	Dictionary. Grammar. NT
86	Khalkha	Mongolic	Altaic	Mongolia	Literature. Newspapers. Periodicals. Radio. TV. Videos. Dictionary. Grammar. Texts. Bible
87	Khasi	Khasian	Austro-Asiatic	India	Literature. Radio. TV. Dictionary. Grammar. Bible
88	Khmer	Khmer	Austro-Asiatic	Cambogia	Fully developed.
89	Khmu'	Khmuic	Austro-Asiatic	Laos	Radio. TV. Dictionary.

						Grammar. Bible portions
90	Kilivila	Oceanic	Austronesian	Papua Guinea	Nuova	Dictionary. Grammar. Bible Texts.
91	Kiowa	Kiowa-Tanoan	Kiowa-Tanoan	USA		Dictionary. Grammar. Texts.
92	Kiribati	Oceanic	Austronesian	Kiribati		Dictionary. Grammar. Bible Texts.
93	Koasati	Muskogean	Muskogean	USA		Dictionary, Grammar, Bible (uso minimo della scrittura)
94	Kobon	Kalam-Kobon	Trans-New Guinea	Papua Guinea	Nuova	Literature. Grammar. NT
95	Kongo	Bantu	Niger-Congo	Rep. Democratica Congo		Radio. TV. Dictionary. Grammar. Bible
96	Korean	Korean	Korean	Corea del sud		Fully developed.
97	Koromfe	Koromfe	Niger-Congo	Burkina Faso		Literature. Dictionary. Grammar. Texts. NT
98	Koyraboro Senni	Songhay	Songhay	Mali		Literature. Dictionary. Grammar. NT
99	Krongo	Kadugli	Kadu	Sudan		Grammar. Texts. NT
100	Kunama	Kunama	Kunama	Eritrea		Literature. Radio. Videos. Dictionary. Grammar. Texts. Bible
101	Kutenai	Kutenai	Kutenai	Canada		Dictionary. Grammar. Texts
102	Ladakhi	Bodic	Sino-Tibetan	India		Literature. Dictionary. Grammar. Bible
103	Lak	Lak	Nakh-Daghestanian	Russia		Literature. Newspapers. Periodicals. Radio. Videos. Dictionary. Grammar. Bible portions
104	Lakhota	Mississippi Valley Siouan	Siouan	USA		Videos. Dictionary. Grammar. Texts. Bible portions
105	Lango	Western Nilotic	Eastern Sudanic	Uganda		Literature. Newspapers. Radio. Dictionary. Grammar. Texts. Bible
106	Latvian	Baltic	Indo-European	Latvia		Fully developed.
107	Lavukaleve	Lavukaleve	Solomons East Papuan	Solomon Islands		Grammar
108	Lepcha	Himalayish	Sino-Tibetan	India		Radio. Dictionary. Grammar. NT

109	Lezgian	Lezgiç	Nakh-Daghestanian	Russia	Literature. Newspapers. Radio. Dictionary. Grammar. Texts. NT
110	Luvale	Bantu	Niger-Congo	Zambia	Newspapers. Radio. TV. Grammar. Bible
111	Maba	Maban	Maban	Chad	Radio. Dictionary. Grammar. Texts.
112	Makah	Southern Wakashan	Wakashan	USA	Grammar
113	Malagasy	Barito	Austronesian	Madagascar	Fully developed.
114	Mandarin	Chinese	Sino-Tibetan	Cina	Fully developed.
115	Mangarrayi	Mangarrayi	Mangarrayi-Maran	Australia	Dictionary, Grammar, Texts
116	Maori	Oceanic	Austronesian	Nuova Zelanda	Grammar. Bible
117	Mapudungun	Araucanian	Araucanian	Chile	Radio. Dictionary. Grammar. Texts. NT
118	Maranungku	Wagaydy	Western Daly	Australia	Texts
119	Maricopa	Yuman	Hokan	USA	unwritten
120	Marind	Marind-Yaqay	Trans-New Guinea	Indonesia	Dictionary. Grammar
121	Martuthunira	Western Pama-Nyungan	Pama-Nyungan	Australia	Dictionary. Grammar. Texts
122	Maung	Iwaidjan	Iwaidjan	Australia	Dictionary. Grammar. Texts. Bible portions
123	Maybrat	Maybrat	Maybrat	Indonesia	Dictionary. Grammar. NT
124	Meithei	Manipuri	Sino-Tibetan	India	Literature. Radio. Videos. Dictionary. Grammar. Bible
125	Miwok (Southern Sierra)	Miwok	Penutian	USA	Grammar
126	Mixtec (Chalcatongo)	Mixtec	Oto-Manguéan	USA	Dictionary. Grammar. Texts.
127	Mundari	Munda	Austro-Asiatic	Messico	Literature. Dictionary. Grammar. NT
128	Murle	South Surmic	Eastern Sudanic	India	Literature. Periodicals. Radio. Dictionary. Grammar. Texts. Bible
129	Nahuatl (Tetelcingo)	Aztecan	Uto-Aztecan	Sudan del sud	Literature. Dictionary. Grammar. NT
130	Nama	Khoe-Kwadi	Khoe-Kwadi	Namibia	Radio. Dictionary. Grammar. Bible
131	Navajo	Athapaskan	Na-Dene	USA	High literacy. Dictionary. Grammar. Bible
132	Ndyuka	Creoles and Pidgins	other	Suriname	Literature. Radio. Dictionary. Grammar. Texts. NT

133	Nenets (Forest)	Samoyedic		Uralic	Russia	Literature. Radio. TV. Dictionary. Grammar. Texts. Bible portions
134	Nez Perce	Sahaptian		Penutian	USA	Dictionary. Grammar. Texts. Bible portions
135	Ngiti	Lendu		Central Sudanic	Repubblica democratica del Congo	Grammar, NT
136	Ngiyambaa	Southeastern Pama-Nyungan		Pama-Nyungan	Australia	Grammar
137	Nivkh	Nivkh		Nivkh	Russia	Dictionary, Grammar
138	Nkore-Kiga	Bantu		Niger-Congo	Uganda	Literature. Newspapers. Radio. Dictionary. Grammar. Bible
139	Nubian (Dongolese)	Nubian		Eastern Sudanic	Sudan	Dictionary, Grammar
140	Nunggubuyu	Nunggubuyu		Gunwinyguan	Australia	Dictionary. Grammar. Texts. NT
141	Oneida	Northern Iroquoian		Iroquoian	Canada	Dictionary. Grammar. Bible
142	Oromo (Harar)	Lowland Cushitic	East	Afro-Asiatic	Etiopia	Dictionary, Grammar, Texts
143	Otomí (Mezquital)	Otomian		Oto-Manguean	Messico	Dictionary. Grammar. NT
144	Paamese	Oceanic		Austronesian	Vanuatu	Dictionary. Grammar. NT
145	Paiwan	Paiwan		Austronesian	Cina-Taiwan	Dictionary. Grammar. Texts. Bible
146	Passamaquoddy-Maliseet	Algonquian		Algic	Canada	Dictionary. Grammar. Bible
147	Paumarí	Arauan		Arauan	Brasile	Dictionary. Grammar. NT
148	Persian	Iranian		Indo-European	Iran	Fully developed.
149	Pirahã	Mura		Mura	Brasile	Grammar, Bible
150	Pitjantjatjara	Western Pama-Nyungan	Pama-	Pama-Nyungan	Australia	Dictionary. Texts. NT
151	Pomo (Southeastern)	Pomoan		Hokan	USA	Grammar
152	Qawasqar	Alacalufan		Alacalufan	Cile	Dictionary, Grammar, Texts, unwritten
153	Quechua (Imbabura)	Quechuan		Quechuan	Ecuador	Dictionary. Texts. Bible
154	Rama	Rama		Chibchan	Nicaragua	Dictionary, Grammar, unwritten
155	Rapanui	Oceanic		Austronesian	Cile	Literature. Radio. Dictionary. Grammar. NT
156	Russian	Slavic		Indo-European	Russia	Fully developed.
157	Sango	Ubangi		Niger-Congo	Repubblica dell'Africa centrale	Radio. TV. Dictionary. Grammar. Bible

158	Sanuma	Yanomam	Yanomam	Venezuela	Grammar. NT
159	Selknam	Chonan Proper	Chonan	Argentina	Dictionary. Grammar
160	Semelai	Aslian	Austro-Asiatic	Malesia	Grammar. Bible
161	Sentani	Sentanic	Sentanic	Indonesia	Dictionary. Grammar. NT
162	Shipibo-Konibo	Panoan	Pano-Tacanan	Peru	Literature. Radio. Dictionary. Grammar. NT
163	Slave	Athapaskan	Na-Dene	Canada	Dictionary. Grammar
164	Spanish	Romance	Indo-European	Spagna	Fully developed.
165	Squamish	Central Salish	Salishan	Canada	Dictionary. Grammar. Texts
166	Suena	Binanderean	Trans-New Guinea	Papua Guinea	Nuova Grammar. NT
167	Supyire	Senufo	Niger-Congo	Mali	Literature. Radio. Videos. Grammar. NT
168	Swahili	Bantu	Niger-Congo	Tanzania	Fully developed.
169	Taba	South Halmahera - West New Guinea	Austronesian	Indonesia	Grammar, unwritten
170	Tagalog	Greater Central Philippine	Austronesian	Filippine	Fully developed.
171	Thai	Kam-Tai	Tai-Kadai	Tailandia	Fully developed.
172	Tiwi	Tiwian	Tiwian	Australia	Dictionary. Texts. Bible
173	Tlingit	Tlingit	Na-Dene	USA	Radio. Dictionary. Grammar. Bible
174	Trumai	Trumai	Trumai	Brasile	No info
175	Tsimshian (Coast)	Tsimshianic	Tsimshianic	Canada	Dictionary. Grammar. Bible
176	Tukang Besi	Celebic	Austronesian	Indonesia	Dictionary, Grammar, unwritten
177	Tunica	Tunica	Tunica	USA	Literature. Dictionary. Grammar. Texts.
178	Turkish	Turkic	Altaic	Turchia	Fully developed.
179	Una	Mek	Trans-New Guinea	Indonesia	Dictionary. NT
180	Ungarinjin	Worrorran	Worrorran	Australia	Dictionary, Grammar, Texts
181	Urubú-Kaapor	Maweti-Guarani	Tupian	Brasile	Dictionary. Grammar. NT
182	Usan	North Adelbert	Trans-New Guinea	Papua Guinea	Nuova Grammar. NT
183	Vietnamese	Vietic	Austro-Asiatic	Vietnam	Fully developed.
184	Wambaya	Wambayan	Mirndi	Australia	Dictionary, Grammar, Texts
185	Warao	Warao	Warao	Venezuela	Dictionary. Grammar. NT
186	Wardaman	Yangmanic	Yangmanic	Australia	Dictionary, Grammar, Texts
187	Wari'	Chapacura- Wanham	Chapacura- Wanham	Brasile	Grammar. Bible

188	Wichí	Matacoan	Matacoan	Argentina	Radio. Bible	Texts.
189	Wichita	Northern Caddoan	Caddoan	USA	Grammar	
190	Yagua	Peba-Yaguan	Peba-Yaguan	Peru	Dictionary. Grammar. NT	
191	Yaqui	Cahita	Uto-Aztecan	Messico	Dictionary. Grammar. NT	
192	Yidiny	Northern Nyungan	Pama-Nyungan	Australia	Dictionary, Grammar, Texts, unwritten	
193	Yimas	Lower Sepik	Ramu-Lower Sepik	Papua Guinea	Nuova Grammar	
194	Yoruba	Defoid	Niger-Congo	Nigeria	Newspapers. Radio. Dictionary. Grammar. Bible	TV.
195	Yuchi	Yuchi	Yuchi	USA	Grammar, unwritten	
196	Yukaghir (Kolyma)	Yukaghir	Yukaghir	Russia	Dictionary, Grammar	
197	Yup'ik (Central)	Eskimo	Eskimo-Aleut	USA	Literature. Dictionary. Grammar. Bible	
198	Yurok	Yurok	Algic	USA	Dictionary. Grammar. Texts	
199	Zoque (Copainalá)	Mixe-Zoque	Mixe-Zoque	Messico	Dictionary. Texts. NT	
200	Zulu	Bantu	Niger-Congo	Sud Africa	Newspapers. Radio. Dictionary. Grammar. Bible	

Ringraziamenti

Vorrei riservare questo spazio finale della mia tesi di laurea ai ringraziamenti verso tutti coloro che hanno contribuito, con il loro instancabile supporto, alla realizzazione della stessa.

Un ringraziamento particolare va alla Professoressa e Relatrice Caterina Mauri, alla mia Correlatrice Francesca Masini e a Ludovica Panetto che mi hanno concesso l'opportunità di esplorare un argomento molto affascinante e mi hanno guidato con pazienza e fiducia, offrendomi preziosi consigli.

Questa tesi rappresenta la chiusura di un percorso durato troppi anni, che non sarebbe stato possibile senza mamma e papà, che hanno sempre creduto in me e hanno avuto la pazienza di aspettarmi e continuare a sostenermi e per questo si meritano il ringraziamento più grande. Quello che sono oggi è anche grazie a voi. Un ringraziamento speciale va anche ai miei nonni per avermi trasmesso i valori importanti della vita. In particolare, a nonna Assunta che ci ama con semplicità, con una torta all'arancia o una lasagna ai carciofi, grazie per essere il collante della nostra famiglia. E a nonno Francesco che mi ha scaldato durante gli inverni freddi con le sue copertine di lana e che non è riuscito a vedermi raggiungere questo traguardo ma io riesco a immaginarmeli i suoi occhi lucidi e felici. Ringrazio mia sorella, che nonostante la distanza geografica, è come se fosse stata sempre qui con me, con le videochiamate per aiuto compiti, eyeliner e fitcheck e per essere la sorella migliore del mondo, e mio fratello, che mi ha mostrato che laurearsi tardi non toglie nulla al proprio valore, ma che anzi è dimostrazione di grande determinazione. Grazie anche a zio Andrea, che ha sempre tifato per me.

Grazie a Biagio per aver sempre creduto in me, per essere la mia persona, per amarci per davvero, per condividere con me le sue passioni, per avermi insegnato la differenza tra un delfino comune e una stenella e per essere sempre la mia fonte di ispirazione.

Vorrei ringraziare anche tutte le persone che hanno condiviso con me, spesso da lontano, con amore e supporto questo viaggio di crescita e apprendimento. Come diceva Thoreau, non esiste nulla che renda il mondo tanto spazioso come avere amici distanti.

A Giorgia, perché avec toi, je suis moi, perché basta uno sguardo per capirci, perché mi ha accompagnato in uno dei periodi più belli della mia vita e per essere, da allora, la sorella che ho scelto.

A Silvia, perché è qui da una vita, perché è lontana ma la sento vicina e per essere riuscite a rimanere noi, sempre uguali ma sempre diverse.

A Nadja e Chiara, per aver condiviso con me gioie e dolori di questi anni universitari.

Ad Adri per i nostri mix linguistici e i nostri discorsi che mi aprono sempre nuove prospettive.

A Francesca, per tutte le nostre chiacchierate mentre tornavamo a casa a Tolosa ma soprattutto per avermi fatto scoprire Napoli e le montanarine.

A Tiziano, per i giochi di parole e le risate a crepelle.

A Dalila, per aver scoperto Bologna e una nuova amicizia insieme.

A Françoise, la mia nonna francese, per avermi fatto sentire a casa e avermi fatto sognare condividendo con me le avventure della sua pazzesca vita.