Rivista Italiana di Dialettologia. Lingue dialetti società a. XLIV (2020), Pendragon Bologna

Marco Angster (Università di Zara), Raffaele Cioffi, Marco Bellante, Livio Gaeta (Università di Torino)

CORPORA E VARIETA' MINORITARIE: LE ISOLE WALSER IN ITALIA

***

Nel contributo vengono considerati i problemi della costruzione di corpora di varietà a bassa densità alla luce di due progetti - DiWaC e ArchiWals - realizzati per preservare il patrimonio linguistico e culturale delle comunità walser di Piemonte e Valle d'Aosta. Nell'articolo si mostra come problemi simili rendano il lavoro su dati parlati e scritti di varietà a bassa densità accidentato. Da un lato le varietà a bassa densità sono per definizione prive o scarse di risorse linguistiche disponibili per la loro elaborazione automatica. D'altro, i dati sia scritti, sia orali di varietà a bassa densità sono caratterizzati da alti livelli di granularità di diverso tipo. Le soluzioni proposte per DiWaC e ArchiWals sono un tentativo di coniugare computabilità e granularità stratificando le informazioni recuperate nei testi originali inclusi nei corpora.

In the paper, the problems of building a corpus of a low-density variety are considered in the light of two projects - DiWaC and ArchiWals - built to preserve the linguistic and cultural heritage of the Walser German communities of Piedmont and Aosta Valley. In the paper it is argued that similar problems affect the task of working on spoken and written data of low-density varieties. On the one hand low-density varieties are defined by the absence or scarcity of ready-to-use language resources for automatic processing. On the other hand, written and spoken data of low-density varieties are both characterised by a high degree of granularity at different levels. The solutions proposed for DiWaC and ArchiWals are an attempt to conjugate computability and granularity by stratifying the information retrieved in the original texts constituting the corpora.

***

0. I progetti Diwac e ArchiWals

Nati con le finalità di studiare, valorizzare e conservare il patrimonio linguistico e letterario delle comunità Walser di area piemontese e valdostana, i progetti

DiWaC e *ArchiWals*[1] costituiscono un tentativo di conciliare un ambito scientifico in notevole espansione (quello della linguistica dei corpora) con i dati di varietà a bassa densità.

In questo lavoro ci riferiremo con la dicitura 'varietà a bassa densità' ad un ampio e variegato spettro di casi. Il termine si ispira a quello, usato nell'ambito della linguistica computazionale, di *low-density languages* (Maxwell, Hughes 2006: 30). Con questo termine si identificano quelle lingue che, a differenza delle poche *high-density languages*, sono povere delle risorse computazionali – innanzitutto testi provvisti di annotazione linguistica – necessarie allo sviluppo di strumenti per il trattamento automatico del linguaggio (NLP). In questa sede preferiamo usare 'varietà' al posto di 'lingue a bassa densità' per includere come entità differenti le varietà, in particolare diamesiche e diatopiche, di lingue (scritte o standard) ad alta e media densità che non necessariamente sono adeguatamente provviste di risorse computazionali. La scelta di 'varietà' ha anche lo scopo di prescindere da discussioni sullo status differente di lingue standard, lingue regionali, dialetti, lingue minoritarie, lingue minacciate, ecc. che a nostro avviso possono in modo trasversale costituire esempi di varietà a bassa densità, secondarie nella prospettiva di questo lavoro.

Utilizzate all'interno di comunità che da decenni sono oggetto di un costante processo di logoramento e invecchiamento, le varietà walser piemontesi e valdostane sono patrimonio linguistico dalle caratteristiche tanto particolari quanto di difficile trattamento in ambito computazionale.

Il patrimonio testuale delle comunità walser, prodotto del revival culturale degli ultimi tre decenni del secolo scorso, appare disomogeneo per tipologia, periodo di stesura e generi. Esito spesso delle iniziative delle singole comunità, tale produzione presenta una notevole varietà di registri, diafasica e diastratica.

Il presente contributo farà perciò riferimento principalmente a dati di scritto, non di parlato. Esso però propone soluzioni che, andando a trattare aspetti come la presenza/assenza di standardizzazione e la granularità del dato linguistico, mostrano una immediata applicazione pratica anche per chi si occupa di corpora che includono sia parlato, sia scritto, o esclusivamente parlato. Inoltre se, almeno fino a questo momento, i corpora dedicati alle varietà walser piemontesi e valdostane non hanno incluso dati di parlato, quanto ideato per il trattamento dello scritto potrebbe avere una diretta applicazione anche a trascrizioni di interviste e di dialoghi.

Nelle varie sezioni del nostro articolo, faremo riferimento in maniera più generale ad alcune caratteristiche del patrimonio a stampa e delle differenti ortografie delle comunità walser. I riferimenti a occorrenze del corpus saranno tratti dalla varietà di Gressoney, che ha rappresentato (insieme a quella di Issime) il primo ambito linguistico del quale il progetto si è occupato in maniera approfondita (Angster *et al.* 2017; Gaeta *et al.* 2019).

1. Corpora di parlato e corpora di lingue piccole o di minoranza

1.1. Corpora e varietà

La maggior diffusione di risorse per il trattamento automatico del linguaggio naturale unita alla generale maggiore diffusione dei computer e di testi in formato digitale sul *web* per un gran numero di lingue[2] ha ormai portato il confine di applicazione della linguistica dei corpora[3] alle soglie delle varietà a bassa densità. Si veda in questo senso il caso delle lingue della ex-Jugoslavia[4], i corpora di tedesco svizzero[5] e austriaco[6]. Non mancano poi esempi di corpora di varietà in contatto[7], di lingue minoritarie[8], di lingue di migrazione[9], anche sotto l'etichetta ormai molto usata di *heritage languages*[10]. Tali risorse rappresentano un ulteriore campo di applicazione della linguistica dei corpora ad ambiti in qualche modo non consueti. Meno vitali, di contro, appaiono tutt'ora le iniziative che prendono in considerazione una o più varietà minacciate: per loro natura scarsamente rappresentate o attestate, spesso in una condizione di obsolescenza e interessate da fenomeni di *language shift*, cioè dall'abbandono della varietà nativa per il codice maggioritario, queste varietà faticano ad avvantaggiarsi delle risorse esistenti in ambito linguistico computazionale. Ciò si deve sia al fatto che dal punto di vista linguistico interno differiscono anche radicalmente da eventuali (e non sempre esistenti) lingue maggioritarie ad esse affini, sia al fatto che le collezioni di testi disponibili sono troppo piccole per essere trattate con gli approcci stocastici tipici della linguistica computazionale, i quali necessitano di una massa critica di dati annotati che in molti casi è più grande degli stessi corpora che per queste varietà si mira a produrre. Tali varietà a bassa densità (o minoritarie) risultano nel contempo di notevole interesse in ambito linguistico e sociolinguistico, in quanto portatrici di fenomeni di tipo sintattico e morfosintattico che le differenziano spesso in maniera sostanziale dalle varietà nazionali di riferimento, scritte o parlate. Non secondaria è poi l'urgenza della documentazione di varietà che nel corso dei prossimi decenni sono con molta probabilità destinate a scomparire.

1.2. Dimensioni e granularità

Abbiamo poco sopra menzionato come per le varietà minacciate siano disponibili, quando esistono, solo corpora di dimensioni contenute. Allo stesso modo, per ragioni differenti, i corpora di parlato tendono anch'essi a raggiungere un numero di token normalmente di molto inferiore, sia alle componenti di lingua scritta nei corpora in cui sono inclusi, sia ai corpora di riferimento per una certa lingua.

Le piccole dimensioni, dunque, tendono ad accomunare i corpora di parlato con quelli di varietà minacciate: ciò non è casuale, da un lato perché spesso queste

varietà sono attestate solo oralmente, per cui i dati disponibili sono trascrizioni di parlato; dall'altro perché, anche nel caso di collezioni di testi scritti, le piccole dimensioni corrispondono frequentemente ad un'alta granularità dei dati stessi.

La granularità dei dati può intendersi in vari modi. Di essa proponiamo qui tre aspetti senza pretesa di esaustività:

– granularità come complessità della trascrizione;
– granularità come complessità della metadatazione;
– granularità come complessità dell'annotazione.

Il primo aspetto della granularità, la trascrizione, riguarda apparentemente soltanto il dato di parlato, ma allargando il concetto di trascrizione a includere quello più ampio di sistema (orto)grafico lo si può applicare anche al caso del dato scritto.

Il secondo aspetto, quello della metadatazione[11], riguarda qualunque tipo di dato, scritto o orale, ma tende ad essere tanto più ricco quanto più il compilatore osserva il dato, per così dire, da una posizione ravvicinata: raccogliendo dati di parlato il ricercatore in molti casi ha un rapporto diretto con il parlante, è coinvolto nella raccolta e tende a provvedere il dato di quante più informazioni possibili riguardanti: parlante, tipologia e contenuto della registrazione, luoghi e tempi della raccolta, ecc. Molte di queste informazioni risultano meno accessibili o del tutto irreperibili nel caso di collezioni di dati raccolte da terzi (acquisizione di registrazioni radiofoniche o televisive; pubblicazioni scritte) oltre a risultare, aumentando la quantità dei dati, meno rilevanti e persino d'intralcio nel caso di analisi quantitative volte ad estrarre generalizzazioni per quanto possibile ampie. Nel caso del vertiginoso crescere della quantità di dati che si osserva con l'emergere dei corpora *web-based* anche una residua metadatazione per generi o tipologie di testo appare ormai quasi superflua, anche perché non vi è più uno schema definito a priori simile a quelli che guidano le raccolte di testi bilanciate, ma dovrebbe essere introdotta a posteriori manualmente.

Il terzo aspetto, quello dell'annotazione, ha conseguenze ambivalenti. Un'annotazione basilare quale quella delle parti del discorso (*POS tagging*, normalmente accompagnato anche dalla lemmatizzazione), se operata su di un corpus ne incrementa vertiginosamente la computabilità e l'arricchimento con altri livelli di annotazione (ad esempio con un *chunking* sintattico). *POS tagging* e lemmatizzazione automatici, però, se facilmente operabili su di un corpus di una lingua o di una varietà ad alta densità (cioè ricca di risorse computazionali), risultano impossibili o scarsamente affidabili se tentati su di una varietà affine la cui (orto) grafia è lontana da quelle standard o maggiormente diffuse. L'alternativa è di norma l'annotazione manuale oppure un nuovo *training* su di un corpus annotato manualmente di dimensioni sufficienti. Tornando dunque al concetto di granula-

rità, l'annotazione ha da un lato alti costi iniziali, ma dall'altro facilita successivi livelli di elaborazione automatica. Dal punto di vista delle dimensioni delle risorse coinvolte, ciascun livello di annotazione può portare ad un loro aumento anche del 100% a partire dalle dimensioni iniziali del corpus puramente testuale.

Il rapporto tra granularità e computabilità è probabilmente l'aspetto più rilevante nel contesto dell'applicazione dei metodi della *corpus linguistics* a dati di parlato o in generale di varietà piccole (cfr. Gaeta in stampa per alcune riflessioni in tal senso).

1.3 Varietà minoritarie e scelte ortografiche

I progetti *DiWaC* e *Archiwals* hanno fra le proprie finalità la valorizzazione e l'analisi delle specificità linguistiche delle varietà walser, sia come sistemi linguistici individuali, sia nel loro insieme. Proprio in ragione dell'oggetto del quale si occupano, tali progetti hanno dovuto affrontare e proporre soluzioni ad alcuni dei problemi generati dal trattamento di lingue a bassa densità. Se per numero di token, infatti, il patrimonio linguistico delle varietà Walser è inevitabilmente ridotto, la complessità e la stratificazione interna del dato appaiono significative. In questo senso, è tanto pregnante la ricchezza dei testi raccolti quanto forte la loro instabilità a livello ortografico. Quest'instabilità, tipica delle varietà minoritarie – dove interessa soprattutto la resa grafica dei fonemi – o in genere delle varietà a bassa densità – dove le questioni di politica linguistica vanno ad influenzare le questioni più strettamente inerenti a grafia e ortografia (Iannàccaro 2010) –, genera una forte variabilità interna e un conseguente basso grado di computabilità del dato.

Ad esempio, la grafia dei testi redatti nella varietà di Gressoney è stata elaborata nei primi decenni del revival culturale walser per opera di alcuni parlanti ed estimatori che in seguito parteciparono alla fondazione dell'associazione culturale *Walser Kulturzentrum* (1982) e alla pubblicazione di varie opere e dizionari delle varietà di Gressoney e della vicina Issime (si vedano WKZ 1988a, 1988b).[12] Queste opere hanno a lungo rappresentato il riferimento – implicito e privo di una standardizzazione cogente – per le successive pratiche scrittorie. La condizione di bassa standardizzazione non è significativamente variata neppure in seguito al tentativo di stabilire una serie di norme grafiche e ortografiche comuni (i risultati si vedano in Antonietti 2010), frutto di uno sforzo congiunto di sportelli linguistici, associazioni culturali e linguisti specialisti di walser, fonetica e grafematica. Se, da un lato, il numero ridotto degli autori ha rappresentato un limite alla proliferazione delle varianti grafiche, il processo di (parziale) stabilizzazione subito dalla grafia ha reso ancora più variegato il quadro generale riscontrabile, non solo fra testi di epoche differenti, ma anche fra documenti appartenenti al medesimo periodo, se non all'interno dello stesso testo.

Un grado di instabilità del dato linguistico del tutto simile, appare comune anche a quelle trascrizioni di parlato (interviste, racconti, interazioni più o meno spontanee tra parlanti) pubblicate nei bollettini curati dalle singole comunità[13]. Problema ancora differente è quello rappresentato dalla trascrizione e dall'acquisizione di quel materiale audio che, accumulatosi nel tempo e registrato nell'ambito di ricerche etnolinguistiche, pur di grande attualità e interesse linguistico, non risulta ancora pubblicato ed è a volte solo parzialmente trascritto.

Le scelte ortografiche e, più in generale, grafiche sono frutto di complesse dinamiche affrontate nella letteratura specialistica di cui non possiamo qui dare conto per ragioni di spazio. Tra i fattori che le influenzano un ruolo non secondario si deve all'incertezza metalinguistica del parlante/scrivente, spesso alfabetizzato con norme ortografiche di lingue standard diverse da quelle minoritarie e non immediatamente utilizzabili per (tra)scrivere la propria varietà.

Nel quadro dell'incertezza della trasposizione nello scritto dei suoni, all'assenza di una norma riconosciuta per le singole varietà si somma una situazione sociolinguistica instabile ed in evoluzione, con il repertorio linguistico della comunità che, per quanto riguarda in particolare le lingue letterarie (o standard) studiate nella comunità, ha visto negli ultimi 150 anni una graduale sostituzione del tedesco con il francese e infine con l'italiano (accanto al francese in virtù dell'autonomia valdostana) come lingue scritte studiate a scuola (cfr. Angster 2014).

Il sistema ortografico vigente, dunque, seppur fondamentalmente basato su di un sistema di corrispondenze analogo al tedesco letterario, risulta in tempi e contesti diversi arricchito di contributi dai sistemi grafici di altre tradizioni. Si vedano ad esempio i tentativi di resa della fricativa palato-alveolare sonora /ʒ/, suono sconosciuto tanto al tedesco quanto all'italiano – se si escludono i prestiti: es. ted. *Jongleur* [ʒɔŋ(g)ˈlø:ɐ̯], oppure italiano *maquillage* /makiˈjaʒ/.

Le autrici di una rubrica pubblicata sui bollettini parrocchiali per più di 20 anni e i cui testi sono stati acquisiti nella nostra banca dati hanno per esempio in un primo momento optato per una scelta 'francese'. Tale scelta, che non ha avuto successo, consisteva nel rappresentare il fonema /ʒ/ – presente in varie parole grammaticali (ad esempio il dimostrativo di prossimità /ˈdɪʒe/ affine al ted. *dieser*) e in vari lessemi (ad esempio nel plurale di "casa", /ˈhiʒer/) – con il grafema <j> (e dunque si trovano *dije*, *hijer*). Più tardi invece la scelta, sancita poi dalla pubblicazione del vocabolario nel 1988 fu quella del *cluster* <sch>, usato già per rappresentare il corrispondente fonema sordo /ʃ/, con l'aggiunta di un diacritico sul segmento <s>: <ŝch>.

Oltre a quella legata all'inventario di corrispondenze tra fonemi e grafemi, gli usi scrittori del titsch di Gressoney presentano una forte variabilità nella rappresentazione di vari fenomeni morfo-fonologici. Un esempio che però forse rappresenta anche la sfida maggiore sia per lo scrivente, sia per il ricercatore, è

la rappresentazione delle sequenze di verbo più pronome enclitico, che nel caso più semplice comportano la presenza di un clitico soltanto – es. *hät es*, lett. 'ha esso (=3SG.N.NOM)' che ricorre rappresentato come *häts* oppure *hätz* –, ma frequentemente presentano anche 2 o 3 elementi clitici a formare un'unica sequenza grafematica, o stringa di caratteri (si veda oltre il § 3.3 per una trattazione più approfondita del problema).

Ulteriori esempi di variabilità possono essere legati a:

– fenomeni morfologico-lessicali: ad es. la scrittura interrotta o meno da spazi nel caso dei verbi con particella separabile: *zròck chéeme*/*zròckchéeme* 'tornare', ma sempre *achéeme* 'arrivare'[14];
– fenomeni sintattici: ad es. l'estensione della grafia *dass* – eventualmente anche *daß*, con la cosiddetta *Eszett*, forma non più corrente nell'attuale norma ortografica (cfr. Gaeta 2017: 88) – che in tedesco standard individua esclusivamente la congiunzione subordinante, anche alla codifica del pronome dimostrativo relativo di genere neutro (*das* in tedesco standard).

A prescindere dal livello di analisi cui i fenomeni si riferiscono, si può intuire quanto frequenti siano i contesti in cui la variabilità ortografica possa avere conseguenze per il trattamento automatico dei dati di una varietà piccola ad ortografia scarsamente standardizzata.

Il quadro qui abbozzato per la varietà di Gressoney (ma si vedano anche Antonietti 2010 e Angster *et al.* 2012 per una panoramica generale sulle comunità walser a sud delle Alpi) dimostra come un primo corposo ostacolo alla computabilità dei dati di questa varietà sia legato ad un tipo di granularità del dato scritto che, seppur radicato in una variabilità ortografica non necessariamente motivata dall'accuratezza della rappresentazione fonetica della lingua, abbia conseguenze molto simili a quelle della granularità di una trascrizione fonetica[15]. In entrambi i casi, infatti, la variabilità nella rappresentazione del segno linguistico ha per conseguenza che i rapporti tra stringa (di testo o fonetica) e lessema siano più complessi di quanto non siano, a livello di corpus, i rapporti tra token e lemma di una varietà scritta standardizzata.

2. Normalizzazione e lemmatizzazione

La computabilità del linguaggio naturale è in gran parte ancora oggi legata all'applicazione di risorse che sono sviluppate sulla lingua scritta e per varietà dotate di una disponibilità elevata di produzione in ortografia standardizzata. Per tale ragione, se si intende sfruttare risorse di NLP esistenti su dati scritti o orali di varietà a bassa densità è necessario rendere il dato da trattare per quanto

possibile vicino, dal punto di vista quanto meno ortografico, alla varietà per cui si dispone di una tale risorsa: bisogna insomma operare una normalizzazione del dato.

In questo senso la stessa trascrizione ortografica del parlato costituisce un passaggio di normalizzazione funzionale al suo successivo trattamento o annotazione (Nagy, Sharma 2014: 236)[16]. La trascrizione ortografica, infatti, permette una riduzione della granularità del dato puramente acustico, ma anche di una eventuale trascrizione fonetica per quanto complessa. In alcuni contesti, inoltre è proprio una trascrizione ortografica a permettere di ottenere in modo automatico una trascrizione fonetica di una registrazione di parlato[17].

Senza pretesa di esaustività, si possono citare come contesti in cui la normalizzazione gioca un ruolo fondamentale almeno i seguenti:

– la comunicazione mediata dal computer (*computer-mediated communication*), in cui i dati possono discostarsi dallo standard per una semplificazione dell'ortografia o per la vicinanza al parlato e l'emergere di tratti linguistici sub-standard o dialettali: si veda Ljubešić *et al.* (2016) per il ripristino dei diacritici in un corpus di lingue slave meridionali estratto da Twitter[18];
– le varietà diacroniche di una stessa lingua, in cui la variazione delle convenzioni ortografiche rispecchia o si somma al cambiamento linguistico: si veda ad esempio Tang *et al.* (2018) per un raffronto tra diverse tecniche automatiche di normalizzazione di varianti ortografiche storiche in corpora di 4 lingue germaniche e ungherese;
– le trascrizioni del parlato di varietà dialettali, in cui la variabilità ortografica rispecchia o si somma alla variazione diatopica: si vedano Samardžić *et al.* (2015) e Scherrer, Ljubešić (2016) per il caso della normalizzazione di trascrizioni di parlato tedesco svizzero nell'ambito del corpus *ArchiMob* (Samardžić *et al.* 2016).

Senza entrare nei dettagli delle varie procedure utilizzate, ci interessa qui sottolineare la dimensione dei *dataset* di allenamento, preparati attraverso procedure automatiche (Ljubešič *et al.* 2016) o manuali:

– Ljubešić *et al.* (2016): oltre 100 milioni di token;
– Tang *et al.* (2018): *dataset* tra 28 mila e 140 mila coppie di token in ortografia storica e token in ortografia standard;
– Scherrer e Ljubešić (2016): 8 mila segmenti prosodici corrispondenti a 65 mila token.

Nei casi in cui le dimensioni totali del corpus o la dispersione dei dati (tra numerose varietà e sottovarietà ciascuna con caratteristiche proprie) rende impratica-

bile il ricorso all'annotazione manuale preliminare di ampi *dataset* di allenamento si devono escogitare soluzioni alternative.

Un esempio di normalizzazione operata manualmente è quello del corpus di SMS in tedesco svizzero – un caso che unisce comunicazione mediata dal computer e variazione diatopica. Ruef, Ueberwasser (2013) descrivono uno strumento per provvedere di glosse interlineari il corpus *sms4science* (http://www.sms4science.ch/ e Dürscheid, Stark 2011): essi propongono un nuovo strumento per ovviare a limiti – come la visualizzazione contemporanea di testo originale e testo glossato/normalizzato e la gestione dei metadati – riscontrati in due strumenti esistenti, ITE (michel.jacobson.free.fr/ITE/) e VARD (http://ucrel.lancs.ac.uk/vard/about/, Baron, Rayson 2008). Interessante è come tali strumenti siano stati sviluppati rispettivamente per l'annotazione di registrazioni audio e per la gestione delle varianti ortografiche in corpora diacronici.

Differente è la soluzione attuata nello sviluppo del corpus diacronico MIDIA (D'Achille, Grossmann 2017). La procedura seguita per annotazione e lemmatizzazione di questo corpus di 8 milioni di token ha sfruttato il modello linguistico sviluppato per *TreeTagger* (Schmid 1994, 1995) per l'annotazione del corpus *web ItWaC* (Baroni *et al.* 2009). Per incrementare i risultati di *TreeTagger*, si è scelto di non produrre nuovi modelli specifici per i vari livelli temporali del corpus e neppure di operare una normalizzazione ortografica preliminare dei testi. La soluzione attuata (Iacobini *et al.* 2014), è stata quella di arricchire il lessico del modello di *TreeTagger* per l'italiano contemporaneo di forme ortograficamente divergenti e desunte da risorse filologiche. Questo approccio ha il vantaggio di evitare l'oneroso processo di annotazione manuale di diversi sotto-corpora di allenamento. Inoltre, arricchendo il lessico del modello per l'italiano esistente si forza il *POS-tagger* a riconoscere durante la lemmatizzazione diverse varianti ortografiche saltando così il passaggio della normalizzazione dei dati originali.

3. Strategie procedurali messe in atto nell'ambito dei progetti *DiWaC* e *ArchiWals*

3.1. Difficoltà di applicare tecniche automatiche di lemmatizzazione su varietà a bassa densità

La mancanza di una standardizzazione grafica rappresenta, come comprensibile, un fattore che si somma alle difficoltà di applicazione di tecniche di lemmatizzazione automatica a varietà a bassa densità. In questo senso, ostacoli significativi sono, nel caso dei corpora a nostra disposizione, sia il limitato numero di token, sia la distanza non trascurabile rispetto alle varietà più prossime e ricche di risorse (il tedesco standard, o lo svizzero tedesco). Un insieme di fattori che

rende difficile anche per i corpora più estesi – di Gressoney e di Formazza – sia l'applicazione dei modelli disponibili di *POS-tagger* (ad esempio l'uso di *TreeTagger* col modello tedesco standard), sia la produzione di un modello specifico per queste varietà: il corpus di addestramento richiesto infatti è di 80000 token circa, pari all'attuale ammontare del corpus di Gressoney e di poco inferiore a quanto acquisito per l'area di Formazza. Tale numero di token supera sia quello attestato per l'area di Issime, sia quello (in verità piuttosto ridotto) di Alagna, sia quello di Rimella. L'annotazione manuale di un corpus di addestramento, inoltre, vista la variabilità ortografica dei corpora stessi, non assicura che il risultato finale sia all'altezza degli sforzi compiuti per ottenerlo[19].

3.2. Soluzioni

3.2.1. Architettura generale della banca dati

Viste le difficoltà di applicazione delle più diffuse tecniche di annotazione automatica, la nostra scelta si è indirizzata verso la progettazione di una nuova piattaforma che potesse adattarsi alla gestione di lingue a bassa densità, costituita da un database multistrato contenente, su due strutture interconnesse, i dizionari prodotti dalle singole comunità da un lato e i corrispondenti corpora testuali dall'altro. Per ciascuna entrata dei dizionari è stata prevista una maschera contenente informazioni lessicali e linguistiche di base (parte del discorso, categorie grammaticali), e campi distinti finalizzati a contenere le tabelle riservate alla flessione nominale, aggettivale e verbale, l'etimologia o le relazioni fra le voci (ad es.: tra composti e derivati e loro rispettive basi, prefissi, particelle, ecc.). Un ulteriore campo è stato poi dedicato alle varianti ortografiche delle singole entrate del dizionario di partenza. Tale struttura di database è stata quindi direttamente interfacciata con una seconda struttura contenente i testi in formato digitale, provvisti di metadati e catalogati per genere letterario e periodo di pubblicazione (Angster *et al.* 2017; Gaeta *et al.* 2019).

La nuova piattaforma è stata pensata per essere flessibile ed adattabile, in futuro, anche ad altre lingue a bassa densità. Ciò che la rende flessibile è proprio la strutturazione su diversi database, ognuno organizzato su più livelli di consistenza del dato (strati). In estrema sintesi, una logica relazionale (SQL) è stata legata a livelli di stringhe e di markup NoSQL, riuscendo quindi ad arricchirli delle qualità tipiche di un database SQL. Ciò permette, ad esempio, di assegnare più valori agli attributi che caratterizzano un singolo elemento (sia questo un lemma a dizionario o una occorrenza del corpus) oppure di ottenere più velocemente i risultati di una query grazie all'indicizzazione istantanea (*on the fly*). Ciò si è rivelato estremamente utile per gestire al meglio la granularità del dato linguistico delle lingue minoritarie da noi prese in esame. La sostanziale

innovazione sta nel fatto di trattare ogni singola entrata lessicale e le relative forme di parola, le singole occorrenze dei corpora (token), ma anche le corrispondenze tra entrate lessicali e occorrenze dei corpora attraverso la logica dei database relazionali. Quindi, ad ogni token viene associato un type collegato ad una serie di metadati che riguardano la forma in questione. Ad ogni entrata lessicale corrisponde un ID univoco, corredato di una serie di attributi specificati nei vari strati del database: varianti grafiche (che costituiscano o meno varianti sociolinguistiche), POS, paradigma (suddiviso a sua volta in elementi singoli derivati e correlati all'ID univoco), traduzioni, ecc. Sul versante corpus la stratificazione del dato consente di immagazzinare i testi in diversi strati, applicando una serie di operazioni automatizzate che servono a classificare ed ottimizzare i testi per l'analisi linguistica[20]. Il testo viene memorizzato in strutture progressivamente incassate l'una nell'altra: innanzitutto come struttura testo effettiva (LONGTEXT) in uno strato; come singole stringhe (STRING, porzioni di testo delimitate da punteggiatura forte) in un secondo strato; in ultima istanza, in quello che è attualmente il terzo strato, come singola occorrenza o posizione di corpus (un token VARCHAR). Anche in questo caso abbiamo un ID univoco a cui vengono correlati tutti i dati degli strati e la metadatazione. Quindi, ad esempio, sarebbe possibile inserire nel corpus un testo e mantenerlo completamente inalterato in modo da essere filologicamente precisi, e nel caso in cui vengano riscontrati evidenti refusi, modificare il testo solamente nelle singole stringhe e di conseguenza avere effetto nei singoli oggetti incassati, ovvero i token. La struttura di questa piattaforma non prevede la necessità di compilare o ricompilare i testi (cioè elaborarli attribuendo un primo livello di marcatura come avviene nei database NoSQL e in XML) per eseguire la lemmatizzazione e il collegamento dei token con le entrate lessicali. Inoltre, è possibile creare infinite sovrastrutture (cioè livelli di annotazione) di elementi sintattici o morfosintattici; ciò rende la struttura adattabile alle caratteristiche di eventuali nuove varietà da analizzare. La struttura e gli algoritmi su cui si basa la piattaforma sono stati depositati come brevetto tecnologico[21].

3.2.2. Collegamento diretto tra lessico e corpus

La piattaforma è stata costruita con la duplice funzione di facilitare la gestione e l'arricchimento del dizionario – costituito di differenti tabelle fra loro interconnesse e contenenti le informazioni lessicali e morfosintattiche – e il suo utilizzo quale base per la lemmatizzazione dei testi del corpus. Infatti, il lessico di base di ciascuna varietà (rappresentato dal dizionario digitalizzato) è stato direttamente interfacciato con il contenuto tokenizzato del corpus, permettendo così una prima lemmatizzazione di massima.

L'alto numero di occorrenze considerate 'nulle', cioè non assegnate automaticamente a nessun lemma, ha messo in evidenza un considerevole numero

di mancate corrispondenze fra lessico acquisito (il dizionario nella sua forma cartacea) e lessico utilizzato nei singoli testi. Dai risultati della lemmatizzazione iniziale si sono potuti riconoscere almeno due ambiti distinti a cui appartenevano le occorrenze non riconosciute: da un lato gli esiti della flessione nominale, aggettivale e verbale; dall'altro il ricco numero di lemmi assenti nel dizionario cartaceo. Catalogate e riordinate, le varie forme riconducibili alle forme flesse sono state collegate ai singoli lemmi di riferimento, e inserite in apposite tabelle che accompagneranno (nella versione successiva della piattaforma) le voci del dizionario. Fra le occorrenze non correttamente lemmatizzabili risulta presente anche un nutrito numero di nomi propri, nomi di luogo, prestiti non acclimatati, onomatopee, espressioni idiomatiche o, in alcuni casi, intere frasi in varietà gallo-romanza. Per ciascuno di essi è stata prevista una collocazione all'interno di un livello a sé stante della piattaforma, così da rendere tali lemmi ricercabili e analizzabili.

In ragione di quanto in precedenza illustrato, risulta comprensibile come una grande parte degli errori di lemmatizzazione causata dal problema delle omografie[22] e da quello delle varianti ortografiche. Tali fattori evidenziano la granularità del dato linguistico e come essa possa gettare luce sui rapporti fra la dimensione orale e la resa scritta della lingua. Per tale ragione, in modo analogo a quanto fatto nell'ambito di MIDIA, invece di normalizzare il testo di partenza si è proceduto alla mappatura delle distinte varianti grafiche, inserendole all'interno di un campo del dizionario[23]. Tale procedimento ha fornito un quadro accurato dei differenti ambiti lessicali e fonologici all'interno dei quali l'oscillazione grafica sembra esercitare un maggiore peso.

Si prenda ad esempio l'oscillazione nella resa palatale o velare di alcuni specifici nessi consonantici (<scht/st> in lemmi come *wòrscht*/*wòrst* o <schp/sp> in *schpäck*/*späck*), così come la forte variabilità nella resa di /ʒ/, rilevabile dalle forme grafiche attribuite al pronome riflessivo *ŝché* (*sché*, *sche*, *je*), o al sostantivo *eŝchél* 'asino' (*eschél* o *ejel*). Di pari interesse, dai dati tratti dall'analisi del corpus, risulta l'oscillazione fra una rappresentazione sorda (fonologica) o sonora (etimologica) del prefisso *p-*/*b-* affine al tedesco *be-* in verbi come *bekennen*, *bedecken*, *behalten*, *besinnen*. Nel titsch di Gressoney occorrono entrambe le varianti benché con varia frequenza: *pchenne*/*bchenne* 'riconoscere' (19 vs. 18 occorrenze); *ptecke*/*btecke* 'coprire' (18 vs. 2 occorrenze); *phoalte*/*bhoalte* 'mantenere' (2 vs. 7 occorrenze); *psénne*/*bsénne* 'ricordarsi' (6 vs. 25 occorrenze). Non secondario, in questo senso, appare come la resa di /ʒ/ sia un fenomeno rilevabile sostanzialmente nei testi del corpus, mentre l'alternanza tra <scht> e <st>, o tra <schp> e <sp>, così come quella delle controparti sorda e sonora del prefisso *p-*/*b-* sia invece rilevabile anche nelle voci presenti nei dizionari prodotti dalle comunità. Tale elemento da un lato appare una conferma della difficoltà nella resa di particolari suoni o fenomeni fonologici da parte degli

scriventi, dall'altro mostra quanto il ricondurre tutte le varianti ortografiche ai rispettivi lemmi possa avvantaggiare la precisione e sistematicità delle ricerche nel corpus.

3.2.3. Istituzione di un livello intermedio tra forma di parola e lemma: le istanze di parola fonologica

Un fenomeno linguistico cui si è accennato sopra in 1.3 e che ha conseguenze rilevanti sulla gestione del rapporto tra lessico e occorrenze del corpus è quello delle sequenze di pronomi enclitici che seguono il verbo di forma finita[24]. Non ci dilungheremo in questa sede su di una descrizione approfondita di questo fenomeno né delle varie interessanti conseguenze che ha a livello teorico. Sia sufficiente per gli scopi della presente trattazione menzionare i seguenti aspetti:

1) il corredo di forme pronominali in queste varietà (come in molte varietà tedesche) include serie toniche e serie atone che possono anche divergere radicalmente tra loro o presentare omofonie estremamente fuorvianti:
   a) 3SG.M/N.DAT: *ém* (forma tonica); *-mò* (forma atona/enclitica; cfr. mat. *imu*);
   b) 3SG.IMPS: *mò* (forma tonica); *-mò* (forma atona/enclitica; cfr. ted. *man*).
2) le forme atone al nominativo seguono direttamente il verbo e precedono quelle all'accusativo o al dativo; tutte possono subire, in aggiunta al fatto di essere forme ridotte, fenomeni di assimilazione che coinvolgono anche le desinenze verbali con la conseguenza che i confini tra i vari elementi componenti la sequenza verbo flesso + clitici divengono spesso fortemente opachi (2.a.) oppure emergono segmenti la cui funzione o origine è difficile da chiarire come nel caso di *dò* in (2.b.), in cui *-d-* verosimilmente risale al suffisso verbale di seconda plurale rianalizzato come parte del pronome *ier* a partire da forme cliticizzate come: *ier heid-ò* > *ier heid-dò* > *heid-er-do*, che troviamo anche in altre combinazioni come *wéntschen-dò* 'vi auguriamo' = augurare.PRES.1PL-2PL.ACC/DAT:
   a) *hämmòne gséd* 'lo si vedeva' < *hät-mò-ne* = avere.PRES.3SG-3SG.IMPS-3SG.M.ACC
   b) *heiderdò erfreit* 'vi siete rallegrati' < *heid-er-dò* = avere.PRES.2PL-2PL.NOM-2PL.ACC/DAT.
3) il venir meno di chiari confini morfofonologici tra desinenze verbali e pronomi enclitici soggetto unito alla tendenza di ripetere la forma tonica prima del verbo flesso è stato interpretato come il primo passo della grammaticalizzazione di nuove desinenze verbali (Giacalone 1989):
   a) *Oanò éndsché sproach wier sibber némme Walser* 'senza la nostra lingua non siamo più walser': *wier sibber* < *wier sind-wier* = 1PL essere.PRES.1PL-1PL.NOM 'noi siamo(-noi)'.

Il quadro così brevemente tratteggiato dovrebbe dare un'idea del problema connesso alla gestione di queste sequenze di verbo e clitici che, come si vede dagli esempi proposti, nella massima parte dei casi vengono trattate dall'ortografia locale come una parola unica. Nella gestione di questo fenomeno dal punto di vista della lemmatizzazione del corpus l'unica soluzione a nostra disposizione, data l'architettura della banca dati come descritta finora, sarebbe stata quella di trattare queste forme come varianti dei rispettivi verbi, eventualmente da integrare in un secondo momento come forme del paradigma verbale. Tuttavia questa soluzione è insoddisfacente per varie ragioni: 1) la difficoltà ai fini della ricerca nel corpus di trovare queste forme che costituiscono invece uno dei fenomeni di interesse linguistico racchiusi nelle produzioni testuali in questa varietà; 2) la scarsa adeguatezza descrittiva di inserire nel paradigma verbale serie di clitici soggetto e complemento.

Abbiamo perciò escogitato una soluzione differente che costituisce un livello aggiuntivo intermedio nella banca dati tra il livello delle occorrenze e quello dei lemmi. Tale livello, che abbiamo denominato "istanze di parola fonologica", ha la funzione, da un lato, di elencare tutte le istanze di questo fenomeno (e non solo, in questa lista sono incluse ad esempio anche le preposizioni articolate) così da poterle facilmente recuperare per l'analisi; dall'altro, ha la funzione di rendere possibile il collegamento tra la specifica occorrenza del corpus e i diversi lemmi e forme di parola che la compongono in modo da accrescere l'adeguatezza descrittiva del sistema di collegamento tra corpus e lessico.

Come per i singoli lemmi a dizionario, per le istanze di parola fonologica è stata raccolta l'estesa gamma di varianti grafiche attestate nel corpus: a ciascun gruppo verbo+clitico è stato attribuito quindi un campo varianti (e.g: *tuemòne*/*tuemone*; *tuemòŝché*/*tuemòsche*/*tuemoje*), compilato parallelamente all'analisi del corpus. Fondamentale per il trattamento di occorrenze del corpus altrimenti non lemmatizzabili in maniera adeguata, la compilazione e il trattamento di tale campo intermedio del database fornisce dunque una casistica in grado di mettere in risalto (in chiave quantitativa e qualitativa) gli effetti sul trattamento grafico delle distinte unità costitutive di tali cluster di quella stessa incertezza del parlante che si osserva nella difficoltà di trasporre il parlato nello scritto secondo una resa grafica stabile.

4. Conclusioni

La gestione e il trattamento di dati linguistici di varietà a bassa densità comportano un differente approccio al dato linguistico. Limitati per numero di token, i corpora di varietà minoritarie presentano di contro molto spesso caratteristiche (una su tutte, l'instabilità ortografica e la sostanziale mancanza di normatività) che inevitabilmente ne vanno a condizionare la computabilità. La notevole granularità del dato è però, di contro, fattore di primaria importanza per un corretto

e completo studio delle differenti caratteristiche morfologiche e fonetiche delle varietà minoritarie, così come dello stretto legame fra la dimensione orale e quella scritta. In questo senso, del tutto necessario appare mettere in atto soluzioni che permettano di valorizzare tali aspetti, e renderli parte integrante del processo di lemmatizzazione e analisi contestuale dei corpora di scritto (così come di parlato) di tali varietà minoritarie. Una soluzione a tale non facile bilanciamento fra necessità di lemmatizzazione e conservazione della varietà interna del dato linguistico è quella di un approccio che (sfruttando le potenzialità del mezzo computazionale) persegua la stratificazione come mezzo per la gestione della granularità del dato su differenti livelli. Ciò concorre alla creazione di uno strumento di studio e di analisi sintattica, morfologica e lessicale di tali varietà isolate.

Note

* Il progetto DiWaC (2015-2017; finanziato dall'Università di Torino in convenzione con la Compagnia di San Paolo: Call 02: Addressing Horizon 2020 – Bando 2014) e il progetto ArchiWals (2017-2020; finanziato dal MIUR con un PRIN – Bando 2015) sono stati entrambi coordinati e diretti da Livio Gaeta. Il presente lavoro è il frutto delle riflessioni sorte nell'ambito di tali progetti. Sebbene la stesura del contributo sia avvenuta per il lavoro congiunto dei vari autori, si attribuisce tuttavia a Marco Angster la responsabilità per le sezioni 1.1, 1.2, 2 e 3.2.3; le sezioni 1.3, 3.1 e 3.2.2 vanno attribuite a Raffaele Cioffi; della stesura della sezione 3.2.1 è responsabile Marco Bellante; dell'introduzione e della conclusione, infine, è responsabile Livio Gaeta.

[1] I risultati dei progetti DiWaC e ArchiWals sono accessibili al sito http://www.archiwals.org/.

[2] Sul recente successo delle iniziative di costruzione di grandi corpora a partire da testi raccolti tramite il *World Wide Web* si vedano Kilgariff, Grefenstette (2003), Baroni *et al.* (2009), Gatto (2014).

[3] La *corpus linguistics* come metodologia emersa nel corso del XX secolo in particolare nella linguistica (grammaticografia e lessicografia) anglosassone attribuisce un significato piuttosto specifico al concetto di corpus, diverso da quello tipico degli studi letterari ("raccolta completa e ordinata di scritti, di uno o più autori, riguardanti una certa materia", De Mauro 2000: *sub voce*). Una definizione di corpus, coerente con i principi della *corpus linguistics* è la seguente: "[A] corpus in modern linguistics, in contrast to being simply any body of text, might more accurately be described as a finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration" (McEnery, Wilson 2006: 32)

[4] Gran parte delle varietà usate nei territori della ex-Jugoslavia dispongono oggi di corpora *web-based* di dimensioni paragonabili ai corpora *web* di lingue come inglese, tedesco, italiano (si vedano Ljubešić, Erjavec 2011 e Ljubešić, Klubička 2014).

[5] In merito, *ArchiMob Corpus* (https://www.spur.uzh.ch/en/departments/research/textgroup/ArchiMob.html).

[6] Fra questi, si ricordano: GRASS (*Graz corpus of Read and Spontaneous Speech*; Schuppler *et al.* 2014) e *Austrian Media Corpus* (https://www.oeaw.ac.at/acdh/tools/amc-austria-media-corpus/; Jutta *et al.* 2013).

[7] Si veda il corpus multilingue Kontatto (https://kontatti.projects.unibz.it/before-kontatti/; Dal Negro, Ciccolone 2018) in cui è documentato il contatto tra italiano e tedesco nella Bassa Atesina.

[8] Per il bretone, si veda il progetto *Breton Text Corpora* accessibile dalla piattaforma *SketchEngine* (https://www.sketchengine.eu/corpora-and-languages/breton-text-corpora/), e il corpus contenuto nella *Leipzig Corpora Collection* (https://corpora.uni-leipzig.de/en?corpusId=bre_wikipedia_2007).

[9] Si veda il progetto *The roots of ethnolects. An experimental comparative study* (Hinskens 2011) incentrato sull'emergere di due etnoletti tra giovani di discendenza turca e marocchina nelle città olandesi di Amsterdam e Nimega.

[10] Si veda il corpus HerLD (*Heritage Language Documentation Corpus*; Nagy 2017) – nato nell'ambito del progetto *Heritage Language Variation and Change in Toronto* (http://projects.chass.utoronto.ca/ngn/HLVC/0_0_home.php) – che raccoglie conversazioni in 10 diverse 'lingue patrimoniali' (*heritage languages*), tra cui l'italiano e il francoprovenzale di Faeto.

[11] Per metadatazione si intende l'assegnazione di dati descrittivi a documenti e file caricati in un archivio. I metadati sono dunque un sistema semantico che fornisce un inquadramento del contenuto di un documento (metadati descrittivi e strutturali) e del contesto nel quale esso si inserisce (metadati gestionali). Tale insieme di informazioni permette una più immediata organizzazione dei documenti, una più veloce ricerca delle informazioni al loro interno e una interoperabilità fra sistemi di gestione e archivi (si veda quanto proposto nell'ambito del *Dublin Core Metadata Initiative*: https://dublincore.org/).

[12] Non va dimenticato, in questo contesto, Zürrer (1982), opera sulla varietà di Gressoney con un'ampia introduzione storica e sociolinguistica seguita da una descrizione essenziale di fonologia e morfologia flessiva. La trascrizione fonetica di stampo dialettologico usata in quest'opera, sistematica e fondata scientificamente, pur essendo facilmente adattabile agli usi scrittori, è stata di fatto ignorata per la stesura del vocabolario di Gressoney.

[13] Di un certo interesse appare il caso della documentazione issimese che conserva lunghe trascrizioni di dialoghi e interviste condotte negli anni dai redattori della rivista *Augusta*. Oggetto di successiva rielaborazione da parte degli stessi redattori della rivista, il testo di questi dialoghi presenta una evidente normalizzazione grafica che, però, non impedisce di percepire la forte varietà ortografica interna, così come l'evidente influsso dell'italiano e dei dialetti romanzi nella morfologia e nel lessico dell'issimese.

[14] Com'è noto, la separabilità e la sua codifica grafica sono problematiche anche nella varietà standard di tedesco: si veda la breve discussione in Gaeta (2017: 85) e più estesamente Fuhrhop (2007).

[15] Ci si concentrerà qui in maniera pressoché esclusiva sulla necessità di fornire una corretta computabilità del dato linguistico. Naturalmente, nella costruzione e gestione degli archivi linguistici e testuali delle varietà Walser si è tenuto conto anche di cosa rappresentare avendo in mente sia le esigenze dell'utente interessato a fruire del patrimonio linguistico e letterario favorendone il recupero, sia per gli specialisti interessati più specificamente al dato linguistico.

[16] "Within linguistic research, a transcript may be used, for instance, for quantitative analysis of morphosyntactic or discourse variables, as a guide for auditory phonetic analysis, for qualitative analysis of conversation, discourse, or interaction, and for theoretical linguistic analysis" (Nagy, Sharma 2014: 236).

[17] Si veda a questo proposito *WebMaus* (Schiel 1999), che attualmente supporta 23 lingue o varietà linguistiche differenti (Kisler *et al.* 2017). Tale applicazione permette di allineare in maniera automatica registrazioni di parlato con la loro trascrizione testuale, tokenizzando il testo e corredandolo di una trascrizione fonologica in SAMPA. Risultato del processo è l'allineamento forzato del testo di partenza con l'audio corrispondente.

[18] Le divergenze ortografiche e i tratti varietistici possono anche essere utilmente sfruttati: si veda Ljubešić, Kranjčić (2015) per lo sfruttamento di queste deviazioni per discriminare tra varietà vicine e creare corpora specifici per le lingue slave meridionali.

[19] Il numero di token a nostra disposizione appariva troppo limitato anche per l'applicazione di tecniche di riconoscimento automatico e lemmatizzazione (così come di normalizzazione grafica) messe in atto per l'ambito linguistico svizzero-tedesco (Garner *et al.* 2014; Honnet *et al.* 2018; Samardzić *et al.* 2015).

[20] Una struttura simile è sfruttata nei database realizzati da Open Lab e dal CELE per le banche dati ladine BLAD (*Banca Lessicala Ladina*) e VOLF e per il VSI (*Vocabolario dei dialetti della Svizzera Italiana*). Ringraziamo un revisore anonimo per la segnalazione.

[21] Brevetto depositato con numero di priorità 102019000021837 in data 21/11/2019.

[22] Le omografie sono state progressivamente controllate e annotate manualmente, correggendo le occorrenze che erano state attribuite dal sistema (sui base probabilistica) in maniera erronea.

[23] Una prima catalogazione delle varianti grafiche era rilevabile nella versione cartacea dei dizionari: alcune delle voci presentavano già chiaramente indicate delle varianti, anch'esse inserite nel database come parte del lessico comunitario.

[24] Nel corpus si incontrano anche parecchi casi di elementi proclitici: tra tutti si può citare ad esempio *z'* che può valere come articolo o come preposizione. I proclitici sono stati in genere trattati come token separati e ricondotti ai rispettivi lemmi.

Riferimenti bibliografici

Angster, Marco 2014, "Lingue di minoranza e di maggioranza. 200 anni di lingue straniere a Gressoney (AO)". In: Valentina Porcellana, Federica Diémoz (a cura di) 2014, *Minoranze in mutamento: Etnicità, lingue e processi etnografici nelle valli alpine italiane*. Alessandria, Dell'Orso: 105-121.

Angster, Marco, Marco Bellante, Raffaele Cioffi, Livio Gaeta 2017, "I progetti DiWaC e ArchiWals". In: Livio Gaeta (a cura di) 2017, *Le isole linguistiche tedescofone in Italia: la situazione attuale e le prospettive future (Workshop, Torino 24 febbraio 2017)* (= *Bollettino dell'Atlante Linguistico Italiano* 41): 83-94.

Angster, Marco, Matteo Rivoira, Antonio Romano 2012, "Eredità, sviluppo interno e contatto. Tratti fonetici, marche morfologiche e scelte (orto)grafiche per le comunità walser di Piemonte e Valle d'Aosta". In: Tullio Telmon, Gianmario Raimondi, Luisa Revelli (a cura di), *Coesistenze linguistiche nell'Italia pre- e postunitaria (Atti del XLV Convegno della Società di Linguistica Italiana (Aosta-Bard-Torino 26-28 settembre 2011)*. Roma, Bulzoni: 331-346.

Antonietti, Federica (a cura di) 2010, *Scrivere tra i Walser. Per un'ortografia delle parlate alemanniche in Italia*. Formazza, Associazione Walser Formazza.

Baron, Alistair, Paul Rayson 2008, "VARD 2: A tool for dealing with spelling variation in historical corpora". In: *Proceedings of the Postgraduate Conference in Corpus Linguistics*. Birmingham, Aston University.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi, Eros Zanchetta 2009, "The WaCky wide web: a collection of very large linguistically processed web-crawled corpora". *Language Resources and Evaluation* 43/3: 209-226.

D'Achille, Paolo, Maria Grossmann 2017, *Per la storia della formazione delle parole in italiano. Un nuovo corpus in rete (MIDIA) e nuove prospettive di studio*. Firenze, Cesati.

Dal Negro, Silvia, Simone Ciccolone 2018. "Il parlato bilingue: italiano e tedesco a contatto in un corpus sudtirolese". In: Felisa Bermejo Calleja, Peggy Katelhön (a cura di), *Lingua parlata. Un confronto fra l'italiano e alcune lingue europee.* Berlin, Peter Lang: 385-407.

De Mauro, Tullio 2000, *Il Nuovo De Mauro*. https://dizionario.internazionale.it/ (ultimo accesso: 25 maggio 2020).

Dürscheid, Christa, Elisabeth Stark 2011, "*sms4science*: An International Corpus-Based Texting Project and the Specific Challenges for Multilingual Switzerland". In: Crispin Thurlow, Kristine Mroczek (eds), *Digital Discourse: Language in the New Media*. New York-London, Oxford University Press: 299-320.

Fuhrhop, Nanna 2007, *Zwischen Wort und Syntagma*. Tübingen, Niemeyer.

Gaeta, Livio 2017, *Lineamenti di grammatica tedesca*. Roma, Carocci.

Gaeta, Livio in stampa, "The Observer's Paradox meets Corpus Linguistics: Written and oral sources for the Walser linguistic islands in Italy". In: Monica Genesin, Gerhard Hempel, Thede Kahl (eds), *Endangered linguistic varieties and minorities in Italy and the Balkans*. Wien, Austrian Academy of Sciences.

Gaeta, Livio (a cura di) 2017, *Le isole linguistiche tedescofone in Italia: la situazione attuale e le prospettive future (Workshop, Torino 24 febbraio 2017)* (= *Bollettino dell'Atlante Linguistico Italiano* 41).

Gaeta, Livio, Marco Angster, Marco Bellante, Raffaele Cioffi 2019, "Conservazione e innovazione nelle varietà Walser: i progetti DiWaC e ArchiWals". In: Roberto Rosselli Del Turco (a cura di), *Dall'Indeuropeo al Germanico: problemi di linguistica storica (atti del XVIII Seminario avanzato in Filologia Germanica, Torino, 18-20 settembre 2017)*. Alessandria, Dell'Orso: 141-193.

Garner, Philip N., David Imseng, Thomas Meyer 2014, "Automatic speech recognition and translation of a Swiss German dialect: Walliserdeutsch". In: *INTERSPEECH-2014*: 2118-2122.

Gatto, Maristella 2014, *The Web As Corpus. Theory and practice*. London-New York, Bloomsbury.

Giacalone Ramat, Anna 1989, "Per una caratterizzazione linguistica e sociolinguistica dell'area Walser". In: Enrico Rizzi (a cura di) 1989, *Lingua e comunicazione simbolica nella cultura Walser. Atti del VI convegno internazionale di studi Walser*. Anzola d'Ossola, Fondazione E. Monti: 37-66.

Hinskens, Frans 2011, "Emerging Moroccan and Turkish varieties of Dutch: Ethnolects or ethnic styles?". In: Friederike Kern, Margret Selting (a cura di), *Ethnic Styles of Speaking in European Metropolitan Areas*. Amsterdam, John Benjamins: 102-131.

Honnet, Pierre Edouard, Andrei Popescu-Belis, Claudiu Musat, Michael Baeriswyl 2018, "Machine Translation of Low-Resource Spoken Dialects: Strategies for Normalizing Swiss German", In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*: 3781-3788.

Hundt, Marianne, Nadja Nesselhauf, Carolin Biewer (eds) 2007, *Corpus Linguistics and the Web*. Amsterdam, Rodopi.

Iacobini, Claudio, Aurelio De Rosa, Giovanna Schirato 2014, "Part-of-Speech tagging strategy for MIDIA: a diachronic corpus of the Italian language". In: *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & of the Fourth International Workshop EVALITA 2014 : 9-11 December 2014, Pisa*: 213-218.

Iannàccaro, Gabriele 2010, "Vita comunitaria e pianificazione linguistica: i Walser". In: Federica Antonietti (a cura di), *Scrivere tra i Walser. Per un'ortografia delle parlate alemanniche in Italia*. Formazza, Associazione Walser Formazza: 15-28.

Jutta, Ransmayr, Karlheinz Mörth, Matej Ďurčo 2013, "Linguistic Variation In The Austrian Media Corpus. Dealing With The Challenges Of Large Amounts Of Data". *Procedia - Social and Behavioral Sciences* 95: 111-115.

Kilgarriff, Adam, Gregory Grefenstette 2003, "Introduction to the Special Issue on the Web as Corpus". *Computational Linguistics* 29/3: 333-47.

Kisler, Thomas, Uwe Reichel, Florian Schiel 2017, "Multilingual processing of speech via web services". *Computer Speech & Language* 45: 326-347.

Ljubešić, Nikola, Tomaž Erjavec 2011, "hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene". In: Ivan Habernal, Václav Matousek (eds), *Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings*. Berlin, Springer: 395-402.

Ljubešić, Nikola, Filip Klubička 2014, "{bs,hr,sr}WaC – Web corpora of Bosnian, Croatian and Serbian". In: Felix Bildhauer, Roland Schäfer (eds), *Proceedings of the 9th Web as Corpus Workshop (WaC-9) @ EACL 2014*. Association for Computational Linguistics: 29-35.

Ljubešić, Nikola, Tomaž Erjavec, Darja Fišer 2016, "Corpus-Based Diacritic Restoration for South Slavic Languages". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*: 3612-3616.

Ljubešić, Nikola, Denis Kranjčić 2015, "Discriminating Between Closely Related Languages on Twitter". *Informatica* 39: 1-8.

Maxwell, Mike, Baden Hughes 2006, "Frontiers in Linguistic Annotation for Lower-Density Languages". In: *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006*. Association for Computational Linguistics: 29-37.

McEnery, Tony, Andrew Wilson 2006, *Corpus Linguistics. An Introduction. Second Edition*. Edinburgh, Edinburgh University Press.

Nagy, Naomi 2017, "Documenting variation in (endangered) heritage languages: How and why". In: Kristine A. Hildebrandt, Carmen Jany, Wilson Silva (eds), *Documenting Variation in Endangered Languages*. Honolulu, University of Hawai'i Press: 33-64.

Nagy, Naomi, Devyani Sharma 2013, "Transcription". In: Robert J. Podesva, Devyani Sharma (eds), *Research Methods in Linguistics*. Cambridge, Cambridge University Press: 235-256.

Ruef Beni, Simone Ueberwasser 2013, "The Taming of a Dialect: Interlinear Glossing of Swiss German Text Messages". In: Marcos Zampieri, Sascha Diwersy (eds), *Non-standard Data Sources in Corpus-based Research*. Aachen, Shaker: 61-68.

Samardžic, Tania, Yves Scherrer, Elvira Glaser 2015, "Normalising orthographic and dialectal variants for the automatic processing of Swiss German". In: *Proceedings of the 7th Language and Technology Conference. Poznan*.

Samardžić, Tanja, Yves Scherrer, Elvira Glaser 2016, "ArchiMob – a corpus of spoken Swiss German". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*: 4061-4066.

Scherrer Yves, Nikola Ljubešić 2016, "Automatic Normalisation of the Swiss German ArchiMob corpus using character-level machine translation". In: *Proceedings of the 13th Conference on Natural Language Processing (KONVENS) 2016*: 248-255.

Schmid, Helmut 1994, "Probabilistic Part-of-Speech Tagging Using Decision Trees". In: *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Schmid, Helmut 1995, "Improvements in Part-of-Speech Tagging with an Application to German". In: *Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland*: 47-50.

Schiel, Florian 1999, "Automatic Phonetic Transcription of Non-Prompted Speech". In: *Proceedings of the ICPhS 1999. San Francisco, August 1999*: 607-610.

Schuppler, Barbara, Martin Hagmueller, Juan A. Morales-Cordovilla, Hannes Pessentheiner 2014, "GRASS: the Graz corpus of Read And Spontaneous Speech". In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*: 1465-1470.

Tang, Gongbo, Fabienne Cap, Eva Petterson, Joakim Nivre 2018, "An evaluation of neural machine translation models on historical spelling normalization". In: *Proceedings of the 27th International Conference on Computational Linguistics*: 1320-1331.

WKZ 1988a: Centro Studi e Cultura Walser / Walser Kulturzentrum. (ed.) 1988, *Greschòneytitsch. Italiano / Titsch*. Quart, Musumeci.

WKZ 1988b: Centro Studi e Cultura Walser / Walser Kulturzentrum. (ed.) 1988, *D'Èischemetöitschu. Italiano / Töitschu*. Quart, Musumeci.

Zürrer, Peter 1982, *Wörterbuch der Mundart von Gressoney. Mit einer Einführung in die Sprachsituation und einem grammatischen Abriß*. Frauenfeld, Huber.