DOI: 10.17469/O2106AISV000008

GRAZIANO TISATO

Acquisizione dell'intero AIS (*Sprach- und Sachatlas Italiens und der Südschweiz*)

The paper presents the results of the NavigAIS project, born to acquire the text of the whole AIS, the Linguistic and Ethnographic Atlas of Italy and Southern Switzerland (*Sprach- und Sachatlas Italiens und der Südschweiz*) (Karl Jaberg and Jakob Jud, 1928-1940). The project was developed between 2009 and 2019 in three complementary phases: 1) the realization in 2009 of NavigAIS, a high resolution digital version of the AIS atlas, provided with a search and navigation engine, to allow the exploration of the 1705 maps contained in the atlas (www3.pd.istc.cnr.it/navigais). NavigAIS assured the real time access to the AIS maps to check on the field the speaker's answers, during the recording sessions of AMDV (the Multimedia Atlas of Veneto Dialects) (www.pd.istc.cnr.it/amdv) in 2009-2010, and to implement the database of the AMDV lemma and the related AIS lemma collected in 1921 in the Veneto region; 2) the NavigAIS online version in 2014 (www3.pd.istc.cnr.it/navigais-web), which does not require a software installation; 3) In 2016, within the *AIS Reloaded* project (www.rose.uzh.ch/de/forschung/forschungamrose/projekte/AIS-reloaded.html), NavigAIS has been provided with a specific OCR, to acquire the AIS text in acceptable times. The estimate is to complete 50% of the AIS tables by the end of 2019.

Key words: linguistic geography, AIS atlas, ancient document acquisition.

1. *Introduzione*

Si presenta in questo lavoro il progetto NavigAIS di acquisizione digitale del testo dell'intero AIS, l'Atlante linguistico ed etnografico dell'Italia e della Svizzera meridionale (*Sprach- und Sachatlas Italiens und der Südschweiz*) (Jaberg, Jud, 1928-1940), e si discutono le problematiche e le soluzioni adottate nel software implementato a questo scopo fra il 2009 e 2019.

Si illustrano le tre fasi in cui si è sviluppato il progetto NavigAIS:

1. La versione digitale e navigabile dell'AIS, completata nel 2009 (Tisato, 2010) www3.pd.istc.cnr.it/navigais, per la realizzazione dell'Atlante Multimediale dei Dialetti Veneti (AMDV) www.pd.istc.cnr.it/amdv.
2. La versione online, consultabile in rete senza necessità di installazione) realizzata nel 2014, www3.pd.istc.cnr.it/navigais-web.
3. La versione con OCR (*Optical Character Recognition*), implementata nel 2015-2016 per il progetto *AIS Reloaded* (Michele Loporcaro e Stephan Schmid, Università di Zurigo, finanziato dal SNSF *Swiss National Science Foundation*, https://www.rose.uzh.ch/de/forschung/forschungamrose/projekte/AIS-reloaded.html; vedi negli atti di questa conferenza: Negrinelli, Donzelli, *Il pro-*

getto AIS reloaded: un archivio sonoro per 36 varietà dialettali della Svizzera meridionale).

Si discutono le caratteristiche del simbolismo AIS, l'architettura e le funzionalità del sistema, le modalità di acquisizione del testo con l'OCR, la validazione dei risultati, la valutazione obiettiva del riconoscitore, ed i tempi di acquisizione dei lemmi.

Si propone, infine, NavigAIS come un modello facilmente adattabile ad altri atlanti linguistici storici. L'implementazione può avvenire in due fasi complementari: una iniziale più rapida, che opera la scansione e il trattamento digitale delle pagine di un atlante, lo indicizza e lo rende navigabile. La seconda forzatamente più lenta, ma più interessante dal punto di vista linguistico, che acquisisce tutto il contenuto testuale dell'atlante con un OCR predisposto al compito. Questo approccio evita che una miniera di informazioni linguistiche ed etnografiche finisca per rimanere sepolta ed inutilizzata negli atlanti cartacei, per i problemi legati alla conservazione dei volumi, alle limitazioni di accesso delle biblioteche, e alle difficoltà di ritrovare l'informazione cercata per mancanza di indicizzazione.

Fig. 1 - *NavigAIS: versione con OCR (NavigAIS-web K. 548.31)*

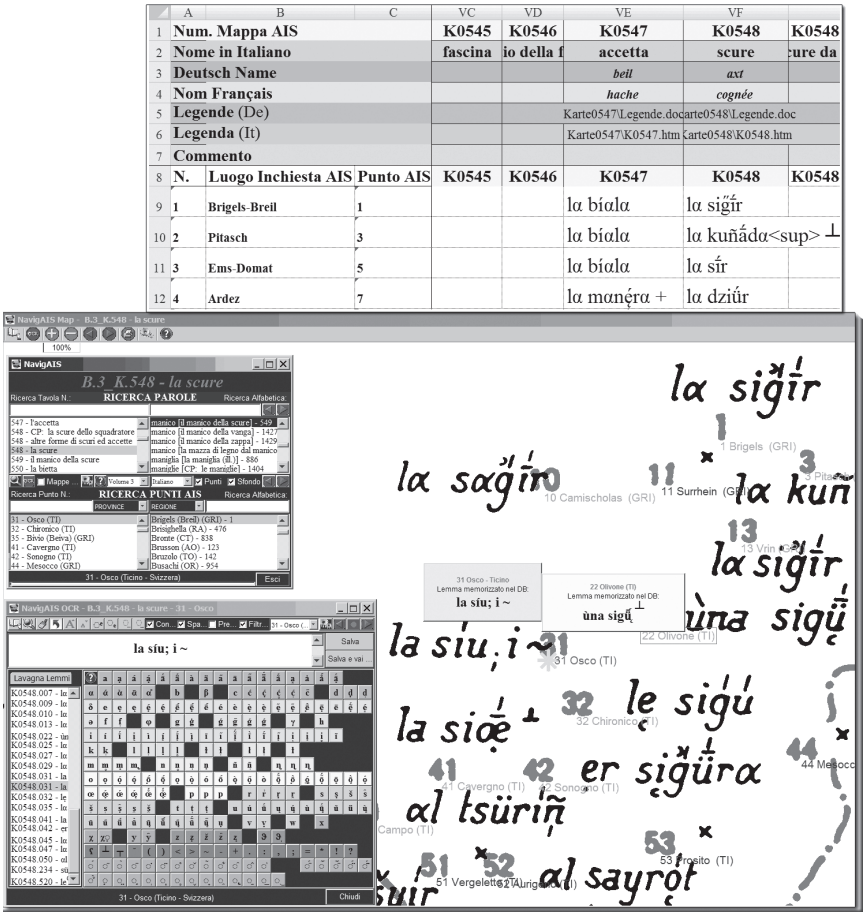
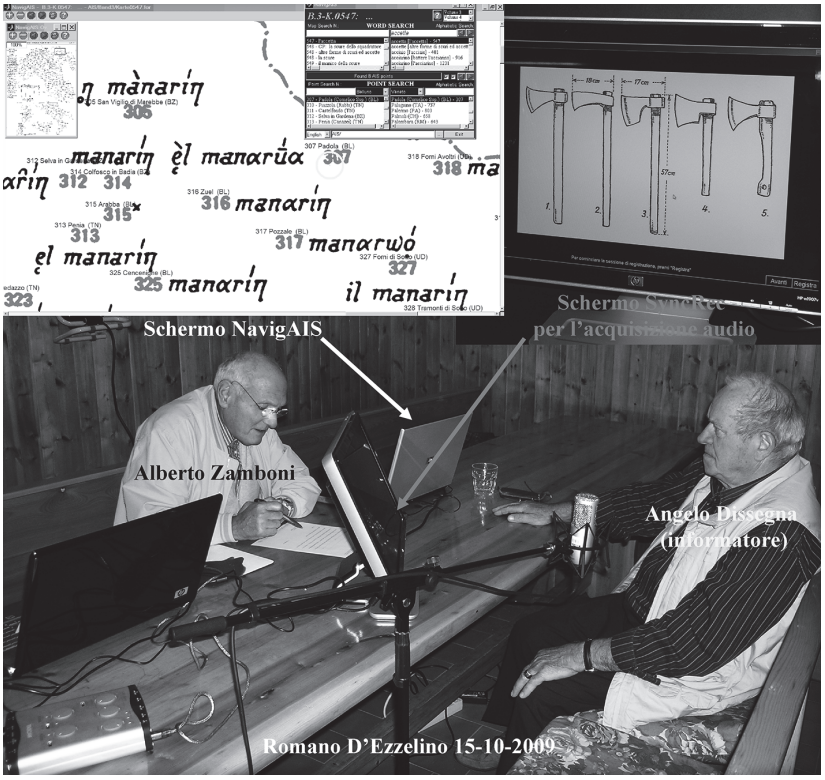


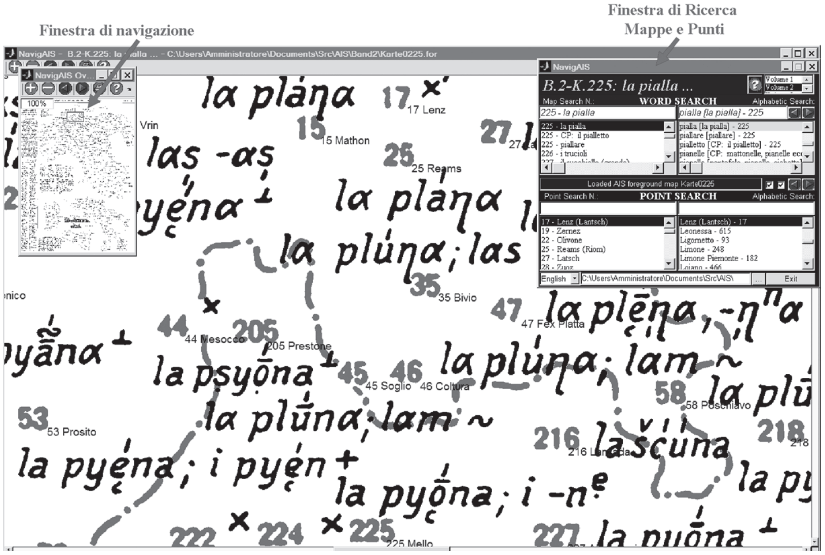
Fig. 2 - *Inchieste AMDV: a sinistra, la postazione del dialettologo, che può confrontare le risposte dell'informatore con quelle AIS sulla mappa di NavigAIS (NavigAIS-web K. 547.307). Il software SyncRec registra l'audio e mostra sequenzialmente l'oggetto corrente all'informatore sullo schermo di fronte a lui, e al responsabile audio sullo schermo nell'angolo sinistro in basso*



2. *NavigAIS Prima Fase: Versione Desktop*

NavigAIS è nato nel 2009 per soddisfare le specifiche esigenze del progetto dell'Atlante Multimediale dei Dialetti Veneti (AMDV, www.pd.istc.cnr.it/amdv) (Tisato, Barbierato, Ferrieri, Gentili & Vigolo, 2013). L'AMDV è stato realizzato fra il 2010 e il 2015 presso il Dipartimento di Discipline Linguistiche, Comunicative e dello Spettacolo dell'Università di Padova, e l'Istituto di Scienze e Tecnologie della Cognizione (ISTC) del CNR di Padova, con un finanziamento della Cassa di Risparmio di Padova e Rovigo per i progetti d'eccellenza del 2008.

Fig. 3 - *NavigAIS: versione desktop (2009) (NavigAIS-web K. 225.35)*



L'idea portante dell'AMDV era il confronto diacronico fra il lessico registrato per l'AMDV nel 2009-2010 e quello raccolto nel Veneto nel 1921 dall'AIS. L'AMDV trae ispirazione da recenti atlanti parlanti: ALEPO (Telmon, Canobbio, 1985), ALD (Goebl, 1994), VIVALDI (Kattenbusch, 1995), ecc., ed in particolare da un precedente lavoro realizzato dall'autore, su scala più ridotta, per i dialetti trentini *Il Trentino dei contadini* (Mott, Kezich & Tisato, 2003). Rispetto all'approccio puramente lessicale dei tradizionali atlanti cartacei, l'AMDV si proponeva di recuperare la dimensione sonora del parlato dialettale e di integrare tutti gli aspetti fonetici, etimologici ed etnografici dei dialetti veneti, in un insieme che permettesse di capire una realtà complessa e articolata come quella dialettale.

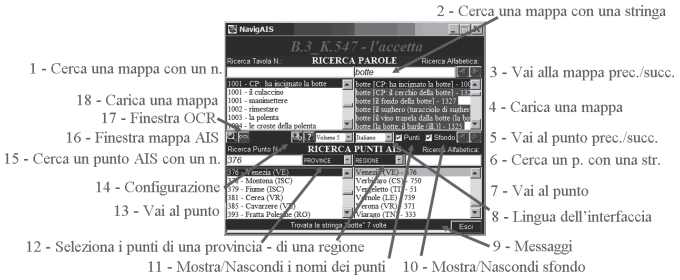
La versione navigabile dell'AIS si rese necessaria non appena divenne evidente che la consultazione dell'atlante non poteva avvenire sui volumi cartacei senza una perdita di tempo enorme, tenuto conto degli orari e delle limitazioni di accesso (feste, ferie, orari, ecc.) della biblioteca del dipartimento universitario, che custodiva i preziosi volumi.

Anche la semplice fotocopia delle pagine AIS era proibitiva, poiché la dimensione fuori del comune (44x58 cm, circa un foglio A2) obbligava a più passaggi per coprire, con un certo margine utile, l'intera superficie, e richiedeva la collaborazione di due persone per girare sottosopra il volume, dargli l'orientamento più appropriato, mantenere la pagina complanare sulla superficie della fotocopiatrice e spostarla poi in un nuovo settore.

Era impossibile pensare di ripetere questa operazione per le 1705 pagine dell'AIS su una normale fotocopiatrice A2 o A3, sia per il problema non trascurabile della conservazione dei volumi originali, sia per la difficoltà operativa di ripescare poi, qualora servisse, una pagina (o una sua parte) in una pila di migliaia di fogli.

La decisione di affrontare alla radice il problema della consultazione dell'Atlante AIS portò nel 2009 a realizzare NavigAIS (Figg. 3, 4), come un progetto completamente autonomo rispetto all'AMDV, quando questo ancora non aveva mosso i primi passi.

Fig. 4 - *Finestra di NavigAIS per la ricerca delle mappe e dei punti AIS. L'esempio mostra i risultati della ricerca della stringa "botte" nell'indice delle mappe AIS*

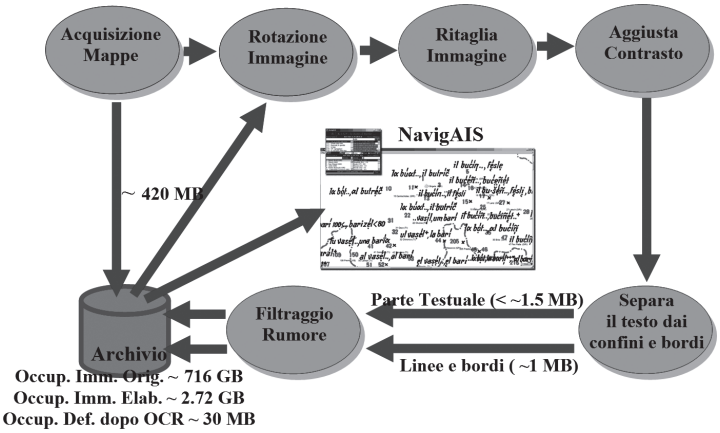


L'operazione preliminare fu la scansione a 600 dpi delle pagine dell'AIS su una macchina professionale messa a disposizione dal Comune di Padova.

L'elaborazione successiva è schematizzabile in 5 fasi (Fig. 5), che possono essere eseguite in un'unica sequenza automatizzata oppure isolatamente, una fase alla volta, sotto il controllo di un supervisore. Al software era richiesto prima di tutto di correggere la rotazione della pagina, dovuta al posizionamento del volume sullo scanner, che non risulta mai quello ottimale. L'aggiustamento della pagina è richiesto per una corretta riproduzione dell'immagine, ma è indispensabile soprattutto per il processo di riconoscimento automatico dei caratteri (OCR), che si prevedeva di attuare in un secondo tempo per acquisire l'intero testo AIS. Dopo la rotazione della pagina, l'immagine era automaticamente ritagliata in corrispondenza dei bordi esterni, in modo da ridurre le dimensioni finali al minimo possibile e l'inclusione di parti della pagina confinante.

Il terzo passo prevedeva un aggiustamento automatico del contrasto dell'immagine con la conseguenza di virare i colori verso le tonalità più scure.

Fig. 5 - *NavigAIS: acquisizione ed elaborazione delle mappe AIS (NavigAIS-web K. 1325.1)*



Il quarto passo serviva a separare il testo dagli altri elementi (bordi, confini regionali, linee orografiche, simboli e numeri indicanti le località dell'Atlante) con un'operazione di sottrazione della parte colorata in arancio, ed a restituire le due mappe complementari della pagina originale. Questo passaggio elimina tutte le componenti di colore inferiori ad una soglia prefissata (tanto per il testo in primo piano quanto per il background), spazzando via completamente lo sfondo rosa pallido ed il rumore di fondo, dovuto, ad esempio, alla comparsa in trasparenza del testo del verso della pagina.

Subito dopo si procedeva ad una ripulitura del "rumore" casuale, e cioè dei punti isolati, con dimensioni inferiori ad una quantità prefissata (il cosiddetto filtraggio *Salt & Pepper*).

L'ultimo stadio salvava le due componenti dell'immagine come maschere binarie fatte unicamente di 0 e 1 ("binarizzazione" dell'immagine originale), per indicare la presenza del colore nero oppure del colore arancio. Questo stratagemma ha permesso di ridurre l'occupazione totale delle 1705 pagine (più le prefazioni) dell'intero Atlante, che in partenza era di 716 GB, a circa 2.72 GB (circa 270 volte inferiore all'occupazione originale).

In base ai dati ottenuti con il progetto AMDV e con il progetto *AIS Reloaded*, la dimensione finale di tutto l'AIS testuale risulterà inferiore a 50 MB.

Una volta ottenute le immagini grafiche delle pagine AIS, è stato messo a punto il software di navigazione in una prima versione *stand-alone* (o *desktop*, che richiede quindi un'installazione sul proprio computer) con l'indicizzazione degli argomenti delle tavole AIS e con un sistema di ricerca delle mappe e dei punti di inchiesta.

Fig. 6 - *AMDV: Ricerca di un tratto fonetico (nasalità) nel database 2009-2010, riscontrato in 1972 lemmi dei 14521 acquisiti in 26 punti di inchiesta del Veneto*

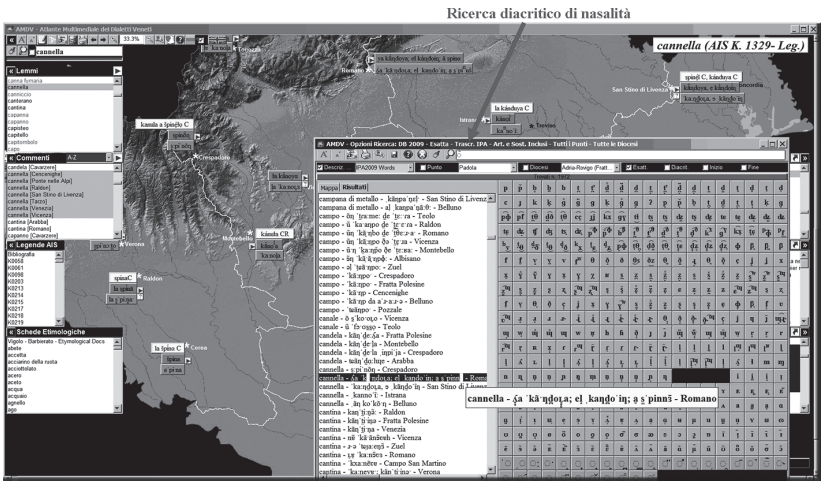
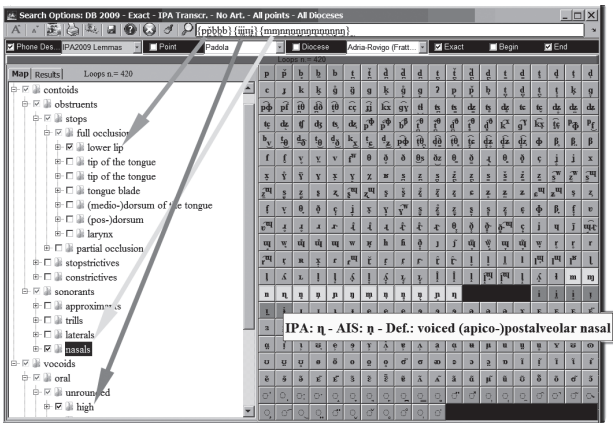


Fig. 7 - *AMDV: Ricerca combinatoria di tutte le occlusive bilabiali, seguite dalle vocali alte e dai contoidi nasali, per un totale di 420 combinazioni distinte*

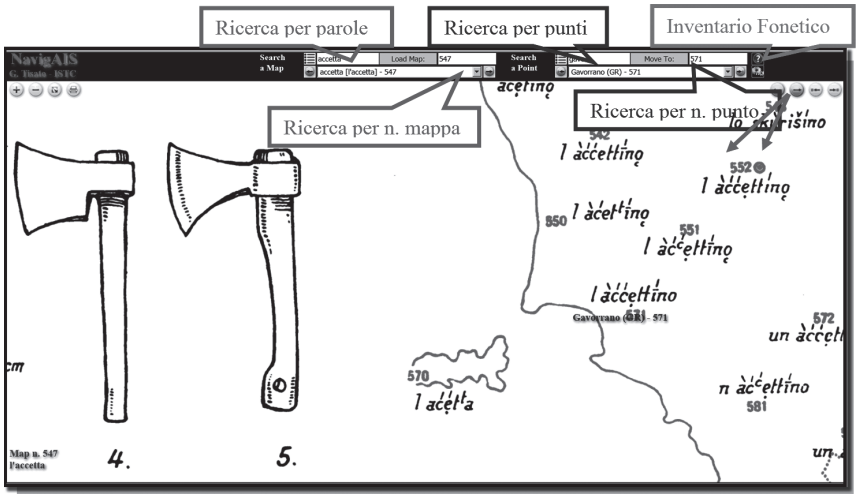


Benché progettato con l'obiettivo finale a lungo termine dell'acquisizione del testo con l'OCR, gli scopi più immediati erano due:

– NavigAIS doveva permettere il confronto in tempo reale sul campo delle risposte degli informatori AMDV con quelle raccolte dall'AIS nel 1921. La Fig. 2 mostra una tipica sessione di lavoro AMDV sul campo: durante l'inchiesta il dialettologo di turno aveva sotto gli occhi le risposte raccolte dall'AIS novant'anni prima, in quello stesso punto e nei punti vicini, e poteva quindi interagire efficacemente con l'informatore dei nostri giorni.
– NavigAIS era anche essenziale per costruire il database delle risposte AIS del 1921 in tempi rapidi e con un minimo tasso di errore. Una volta salvata in un database, la trascrizione corrente poteva essere immediatamente visualizzata fianco a fianco con quella dell'immagine grafica originale, in modo da permettere di correggere gli eventuali errori. Questa strategia ha consentito di realizzare la trascrizione in due mesi di lavoro con un errore inferiore al 2%.

Visto l'interesse che NavigAIS poteva avere per studenti, ricercatori ed esperti della materia, nel gennaio del 2010, l'atlante fu reso disponibile in un sito web: www3.pd.istc.cnr.it/navigais.

Fig. 8 - *NavigAIS-web versione online dell'AIS (2014): le mappe AIS e i punti d'inchiesta sono accessibili con parametri aggiunti al normale indirizzo web (NavigAIS-web K. 547.571)*



3. *NavigAIS Seconda Fase: Versione Online*

La seconda fase, implementata nel 2013-2014, ha permesso di realizzare una versione online dell'atlante AIS (www3.pd.istc.cnr.it/navigais-web) (Fig. 8), che ha il vantaggio di non richiedere una installazione del software ed è consultabile in rete da qualsiasi parte del mondo.

La novità rilevante, rispetto alla versione desktop del 2009, è la possibilità di aprire in rete una certa mappa AIS in un certo punto d'inchiesta, per evidenziare il lemma voluto direttamente dal proprio computer o dalla propria applicazione mediante tre parametri, inseriti nell'indirizzo web:

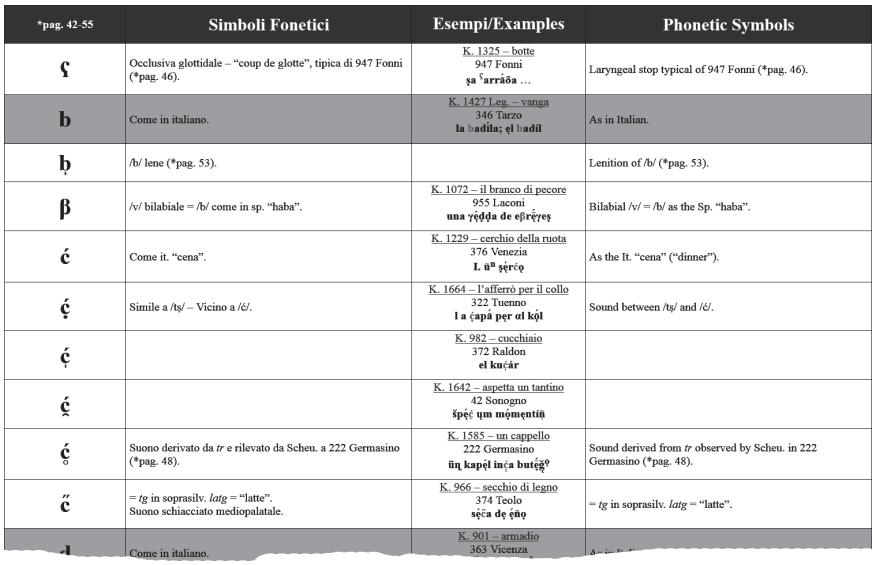
- map = xxxx (xxxx è il numero della tavola AIS 1-1705)
- point = yyy (yyy è il n. identificativo del luogo di inchiesta 1-990)
- loc = luogo (luogo è il nome del luogo di inchiesta)

Si forniscono alcuni esempi concreti di utilizzo dei parametri:

- www3.pd.istc.cnr.it/navigais-web (apre la mappa n. 1 nel punto n. 1, Brigels);
- www3.pd.istc.cnr.it/navigais-web?map=1401 (apre la mappa n. 1401, nel punto n. 1);
- www3.pd.istc.cnr.it/navigais-web?map=1401&loc=Teolo (mappa n. 1401, punto Teolo);
- www3.pd.istc.cnr.it/navigais-web?map=1434&point=229 (apre la mappa n. 1434 nel punto n. 229, corrispondente a Sonico).

La seconda novità è l'inserimento nel software dell'inventario fonetico AIS, tratto dai Cap. 3-4 del volume introduttivo dell'AIS (Jaberg, Jud, 1928), con link diretti ad esempi concreti sulle tavole AIS, sfruttando il meccanismo di indirizzamento appena spiegato, *www3.pd.istc.cnr.it/navigais-web/AIS_symbols.htm*.

Fig. 9 - *NavigAIS-web: Inventario fonetico con esempi (www3.pd.istc.cnr.it/navigais-web/AIS_symbols.htm)*



*pag. 42-55	Simboli Fonetici	Esempi/Examples	Phonetic Symbols
ʔ	Occlusiva glottidale – "coup de glotte", tipica di 947 Fonni (*pag. 46).	K. 1325 – botte 947 Fonni **ṣa ˈarrāða** …	Laryngeal stop typical of 947 Fonni (*pag. 46).
b	Come in italiano.	K. 1427 Leg. – vanga 346 Tarzo **la badìla; el badìl**	As in Italian.
b̥	/b/ lene (*pag. 53).		Lenition of /b/ (*pag. 53).
β	/v/ bilabiale = /b/ come in sp. "haba".	K. 1072 – il branco di pecore 955 Laconi **una ɣéḍḍa de eβɾéɣeṣ**	Bilabial /v/ = /b/ as the Sp. "haba".
ć	Come it. "cena".	K. 1229 – cerchio della ruota 376 Venezia **I. üⁿ ṣérćọ**	As the It. "cena" ("dinner").
ć̣	Simile a /ʦ/ – Vicino a /ć/.	K. 1664 – l'afferrò per il collo 322 Tuenno **l a ć̣apà per ɑl kǫ̀l**	Sound between /ʦ/ and /ć/.
ć̣		K. 982 – cucchiaio 372 Raldon **el kuć̣ár**	
ć̣		K. 1642 – aspetta un tantino 42 Sonogno **špéć̣ ųm mǫ̀mẹntiǫ̃**	
ć̣	Suono derivato da *tr* e rilevato da Scheu. a 222 Germasino (*pag. 48).	K. 1585 – un cappello 222 Germasino **ũn̦ kapẹ̀l inć̣a butéḡǫ̃**	Sound derived from *tr* observed by Scheu. in 222 Germasino (*pag. 48).
c̋	= *tg* in soprasilv. *latg* = "latte". Suono schiacciato mediopalatale.	K. 966 – secchio di legno 374 Teolo **séc̋a dẹ ẹ̃nọ**	= *tg* in soprasilv. *latg* = "latte".
d	Come in italiano.	K. 901 – armadio 363 Vicenza	

4. *NavigAIS Terza Fase: Versione con OCR Integrato*

La terza fase del progetto NavigAIS riguarda l'acquisizione del testo dell'intero Atlante AIS. Questa fase ha lo svantaggio di richiedere tempi di esecuzione lunghi, sia per l'addestramento e l'implementazione dell'OCR, sia per la correzione degli errori del riconoscitore, ma ha d'altra parte il vantaggio enorme di generare un database di dati interrogabili di grande interesse per tutto il campo della dialettologia romanza.

Preliminare a questo passo è la scelta di una rappresentazione dell'informazione che sia adatta alle necessità linguistiche e computazionali successive. Per poter codificare il lessico AIS, che impiega simboli latini e greci ed un numero molto alto di diacritici, non si può ricorrere ad un font specifico (o proprietario), per alcune buone ragioni: *a*) significherebbe ricorrere a diverse centinaia di caratteri diversi l'uno dall'altro (con un aumento delle difficoltà di addestramento dell'OCR); *b*) creerebbe grosse difficoltà computazionali nella ricerca di una stringa di caratteri e difficoltà nel digitare la sequenza corrispondente al carattere voluto; *c*) comporterebbe la necessità di installare il font sulla macchina su cui deve girare il programma applicativo (che richiede a sua volta una fase di installazione), e la conseguente preclusione dell'utilizzo in rete dell'informazione, qualora si volesse una versione online del software.

La soluzione a questi problemi è l'adozione di un font generico, basato sullo standard Unicode, che abbia un insieme di diacritici sufficientemente esteso da coprire tutte le necessità dell'inventario fonetico che si voglia usare. Il Times New Roman, ad esempio, a partire dalla versione 5.0, ha 112 diacritici, sufficienti per

esprimere la grafia AIS abbastanza fedelmente. Il vantaggio di questa soluzione è che permette di implementare l'algoritmo di ricerca di una parola e delle sue varianti in maniera molto più semplice ed efficace, consentendo all'utilizzatore di cercare una sequenza o un tratto fonetico anche senza presupporre la conoscenza a priori della combinazione di tasti per richiamare un simbolo specifico (Figg. 6, 7).

Nell'AMDV, ad esempio, si possono individuare le realizzazioni fonetiche contrassegnate con il diacritico di nasalità [õ] digitandolo nel riquadro di ricerca (come indicato dalla freccia in Fig. 6).

In maniera analoga alla nasalizzazione, si possono ricercare altre caratteristiche come: accento (primario e secondario); durata (crono e semicrono); sollevamento e abbassamento della lingua, avanzamento e arretramento della base della lingua, varie tipologie articolatorie (dentale, alveolare, laminale, ecc.), centralizzazione, arrotondamento, rilascio non udibile, lateralizzazione, palatalizzazione; tipologie di fonazione mista e altre ancora.

Si può inoltre cercare la sequenza all'inizio o alla fine di parola (caselle da spuntare in alto a destra in Fig. 7). Si può restringere la ricerca alle parole riscontrate solo in uno dei punti di inchiesta oppure a quelli appartenenti ad una diocesi particolare (2° e 3° *check-box* in alto in Figg. 6, 7). Si può ottenere la ricerca combinatoria fra vari gruppi fonetici C-V, selezionati dai rami dell'alberatura che compare a sinistra della tastiera virtuale di Fig. 7 (alberatura che fra l'altro mostra le relazioni di parentela fonetica), oppure scrivendo manualmente fra parentesi graffe {} i simboli da combinare nella ricerca. Nell'esempio di Fig. 7, si cercano tutte le combinazioni costituite da un'occlusiva bilabiale, seguita da un vocoide alto e da un contoide nasale (/pim/, /pin/, ecc.), per un totale di 420 (5x6x14) sequenze distinte da individuare.

Per quanto riguarda la trascrizione fonetica dell'Atlante AIS, si decise di attenersi a criteri strettamente filologici e di mantenerla così com'era, senza utilizzare glifi diversi (semplificati), come era stato fatto, per ragioni di leggibilità, nel progetto sui dialetti trentini, e anche senza "tradurla" in IPA, come si sta facendo anche recentemente (ad es., *nell'Atlante Digitale dell'Italia e della Svizzera Meridionale* ADIS; Krefeld, Lücke, 2010[1]).

Si evitava in questo modo di introdurre un secondo livello interpretativo, che si sarebbe aggiunto a quello operato a suo tempo dai raccoglitori dell'AIS (nel caso del Veneto, P. Scheuermeier), con il grave handicap di non disporre più della sorgente sonora originale.

Una controindicazione a questa operazione complessa veniva innanzitutto dal fatto che non esiste una corrispondenza biunivoca fra i simboli dell'inventario fonetico AIS e quello IPA.

La seconda controindicazione era che, inevitabilmente, questo passaggio sfociava nella tentazione di "restaurare" l'AIS, a somiglianza di quanto può avvenire anche in altri campi disciplinari (archeologia, decifrazione di testi antichi ecc.), modificando trascrizioni, giudicate errate, che riportavano invece fedelmente varianti locali significative.

[1] https://www.adis.gwi.uni-muenchen.de/AIS.php.

5. *Scelta di un OCR per l'AIS*

Il compito di riconoscere i caratteri dell'Atlante AIS presenta difficoltà notevoli, da un lato perché sono tracciati a mano, dall'altro per il corsivo e per il numero dei diacritici e dei livelli su cui sono distribuiti, caratteristiche che impediscono l'utilizzo dei normali OCR.

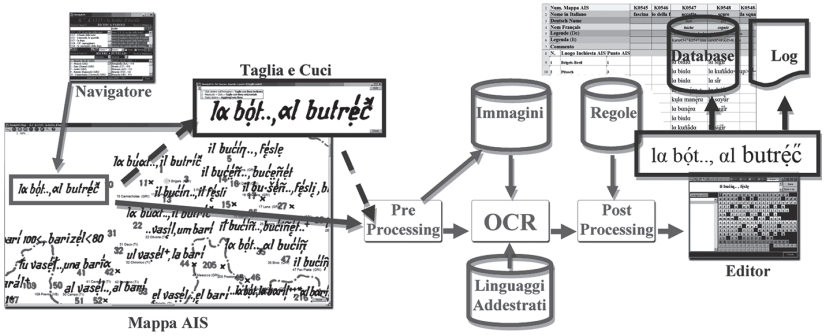
Il riconoscimento dei caratteri a stampa (OCR, *Optical Character Recognition*) e soprattutto dei caratteri scritti a mano (HCR, *Handwritten Character Recognition*) con diacritici ed in corsivo è ancora un campo interdisciplinare aperto ad una grande quantità di ricerche, e allo sviluppo di metodologie e tecnologie innovative (ICR, *Intelligent Character Recognition*, *Machine Learning*, *Pattern Recognition*, *Deep Learning*, ecc.).

Per dare una idea sullo stato dell'arte, si possono citare i risultati di un test, alquanto elementare, come quello di riconoscere una sequenza di 36 caratteri (inglesi, alfa-numerici, maiuscoli, isolati), ripetuta 10 volte, e scritta a mano libera da 7 persone di varia età (Vithlani, Kumbharana, 2015).

La sequenza è stata sottoposta a 6 software OCR sia desktop che online. Il migliore OCR si è dimostrato essere Custom OCR Online, con una percentuale di accuratezza totale del 43.89% di riconoscimento, ovverossia un errore CER (*Character Error Rate*) del 56.11%.

Un risultato così modesto dimostra che siamo ben lontano dalle percentuali vicine al 100% che si ottengono nel caso dei caratteri stampati (quando non siano rovinati o affetti da rumore).

Fig. 10 - *NavigAIS: Schema generale. A sinistra il navigatore. Sulla mappa AIS è indicato con un riquadro la zona dell'immagine passata all'OCR (al centro). A destra, l'Editor per la correzione con la tastiera virtuale e il salvataggio della sequenza riconosciuta*



Da notare che, in genere, le condizioni che si incontrano nel riconoscimento dei caratteri AIS sono peggiori di quelle del test descritto. In generale, si possono considerare le sequenze fonetiche dell'AIS come caratterizzate da un numero molto alto di caratteri a bassa frequenza, con un andamento che può essere descritto dalla legge di Zipf (Zipf, 1949).

La conseguenza inevitabile di questo andamento non può essere altro che il peggioramento dei risultati di un qualsiasi OCR basato su un motore probabilistico. In particolare: *a*) l'atlante AIS non riguarda una lingua sola, ma le lingue di molte regioni con una grande variabilità linguistica. Il numero di simboli da riconoscere aumenta di conseguenza (per es. 230 simboli diversi per il Veneto, ma più di 300 per l'intero AIS) con il conseguente aumento delle probabilità di errore dell'OCR; *b*) il corsivo (circa 70° di inclinazione, vedi Fig. 1-3, 8) dei caratteri AIS abbassa le performance degli odierni OCR a livelli anche inferiori a quelli del testo scritto a mano; *c*) la struttura dei diacritici, articolata su 7 livelli, è al momento irrisolvibile da parte degli attuali OCR (Fig. 9); *d*) manca per l'AIS un modello del linguaggio (o meglio, dei linguaggi), che predica le probabilità della successione di una serie di caratteri e di parole nella(e) lingua(e) da riconoscere, e che permetta all'OCR di scegliere fra varie possibili sequenze quella giusta; *d*) manca per l'AIS un dizionario che elenchi le parole esistenti nei dialetti in questione (forme flesse comprese) e che, come nei correttori ortografici, consenta di individuare la forma corretta fra le varie candidate dall'OCR.

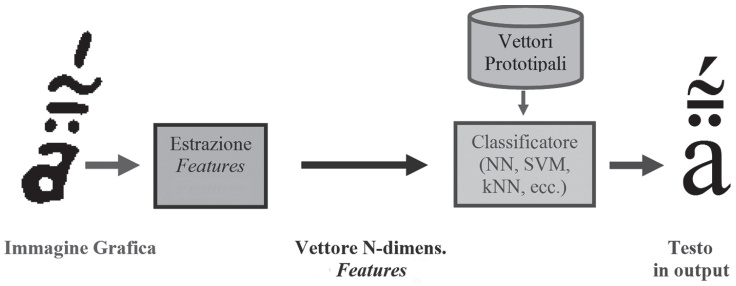
All'inizio della progettazione di NavigAIS furono presi in considerazione tre OCR: i due software commerciali più quotati al momento, Omnipage (Nuance www.nuance.com), e FineReader (ABBYY, www.abbyy.com), ed un software di uso libero Tesseract (sviluppato dalla Hewlett Packard fra il 1985 ed il 1994, diventato open source nel 2005 con licenza Apache 2.0, sponsorizzato da Google per una decina di anni, e ora scaricabile dal sito: https://github.com/tesseract-ocr). Un primo esame rivelò che Omnipage non assicurava una gestione dei diacritici adeguata alle necessità del progetto dell'AIS, e fu scartato.

FineReader permetteva la definizione dei caratteri e dei diacritici AIS secondo la convenzione Unicode. Per quanto riguarda la fase di *training*, il software accettava la ridefinizione delle immagini grafiche dei caratteri problematici con la loro corretta identità, sebbene il meccanismo fosse molto laborioso: in effetti, perché il risultato fosse accettabile, si doveva ripetere l'operazione di identificazione su migliaia di glifi. Da tenere presente anche che questa assegnazione era abbastanza approssimativa, in quanto le immagini catturate da FineReader non potevano essere ripulite dagli spezzoni di linee appartenenti ai caratteri confinanti, con relativi effetti di peggioramento degli errori del riconoscitore.

Dove tuttavia FineReader rivelò la sua inadeguatezza fu nell'impossibilità di integrazione del software nel processo di elaborazione complessivo. In effetti, l'automazione del software consisteva nel depositare da parte dell'utente l'immagine grafica voluta in una cartella Hot Folder, successivamente prelevata e riconosciuta da parte del programma. Purtroppo, l'azione dell'OCR, prevista per funzioni di *Office Automation*, era asincrona e scattava a intervalli temporali discreti con un minimo di 1 minuto. Una simile limitazione comportava tempi di acquisizione inaccettabili per il milione di parole dell'AIS e finiva per rendere la trascrizione manuale più vantaggiosa. A peggiorare le cose si scoprì che il meccanismo di automazione prevedeva solo l'uso di linguaggi standard predefiniti, ed impediva l'uso di un linguaggio addestrato dall'utente, impedendo così l'impiego del software per l'AIS.

Tesseract offriva vantaggi notevoli: 1) metteva a disposizione tutto il codice del programma; 2) poteva essere facilmente integrato nella catena di elaborazione della sequenza fonetica; 3) aveva una percentuale di errore confrontabile con i software commerciali più prestigiosi.

Fig. 11 - *Schema di un generico riconoscitore OCR*



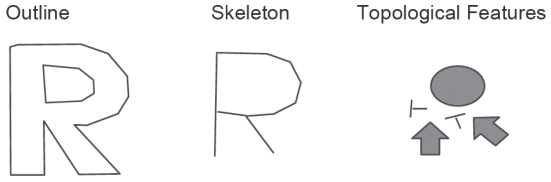
6. *Funzionamento di un OCR*

Il processo di un generico OCR può essere schematizzato in due fasi (Fig. 11):

1. una fase di misurazione, con cui si estraggono dall'immagine grafica in esame un set di misure caratteristiche (*features*) che identificano un glifo;
2. una fase di classificazione con cui si cerca il *matching* fra features ricavati dai caratteri da riconoscere, e prototipi memorizzati nella fase di addestramento (*templates*). Il classificatore deve trovare la minima distanza nello spazio a N dimensioni (con N = numero di features) fra il vettore estratto dal carattere sconosciuto, ed i templates già ricavati dal *training*.

Come è evidente, la scelta del numero di features è la chiave del sistema OCR, poiché porta ad una complessità computazionale ed implementativa, che cresce esponenzialmente con il numero di features. Si tratta di trovare il giusto compromesso fra le esigenze contrastanti della migliore approssimazione possibile e del minor costo computazionale da pagare. I metodi (Fig. 12) più usati sono basati su: 1) il contorno del carattere; 2) la scheletrizzazione del carattere; 3) le proprietà geometriche e topologiche dei caratteri.

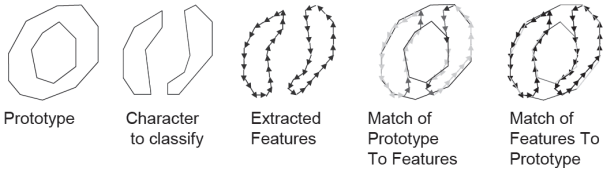
Fig. 12 - *Possibili features da usare nel processo di riconoscimento (Smith, 2014)*



L'approccio che dimostra l'aderenza più stretta ai requisiti ideali di estrazione dei *features* (robustezza, efficienza e peso computazionale) è quello basato su una rappresentazione matematica del contorno dei caratteri.

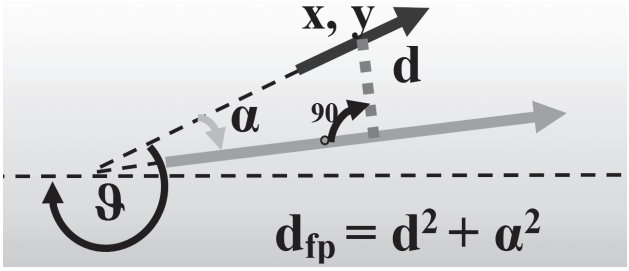
Si evita una descrizione puntuale dell'outline, troppo costosa dal punto di vista computazionale, e si ricorre ad una approssimazione poligonale del contorno del glifo con un numero di lati appropriato, per rendere tutto il processo di riconoscimento sostenibile (Fig. 13).

Fig. 13 - *Approssimazione del contorno di un carattere con una poligonale (linee con le frecce) e matching con features prototipali (linee continue) (Smith, 2014)*



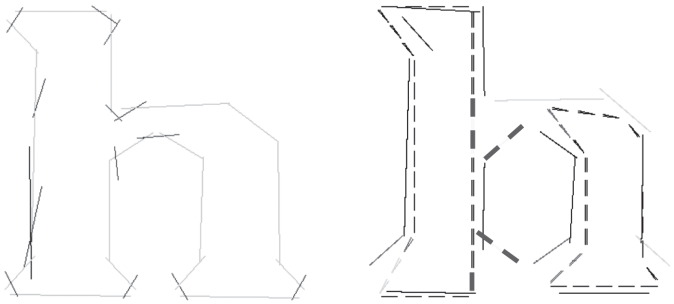
Il contorno è descritto da una serie di vettori 3D (linee in Fig. 14) caratterizzati da tre parametri: la posizione x, y nei vari punti di applicazione del vettore lungo le linee di contorno e l'angolo ϑ che il vettore forma con l'asse orizzontale di base (linea tratteggiata in Fig. 14). Ognuno dei vettori che descrivono il contorno del carattere è confrontato con i vettori prototipali (linea più lunga in Fig. 14) ricavati a priori da un set opportuno di campioni di caratteri della lingua in questione, e memorizzati in un database. La distanza d_{fp} da calcolare per ognuno dei vettori dei modelli di riferimento è data dalla somma della distanza euclidea d (linea tratteggiata in Fig. 14) (perpendicolare al vettore prototipale indicato dalla linea più lunga in Fig. 14) al quadrato e dall'angolo α fra le loro proiezioni al quadrato.

Fig. 14 - *Vettore 3D della poligonale del contorno del carattere, posizionato nel punto x, y e formante l'angolo ϑ con l'asse orizzontale, e distanza d da un feature (Smith, 2014)*



La minima fra le distanze d_{fp} dai prototipi etichettati fornisce l'identità del candidato più probabile del riconoscimento (Fig. 14). Da notare che l'approssimazione poligonale è un'arma a doppio taglio, che ha il vantaggio di eliminare la dentellatura rumorosa irrilevante, ma che ha il difetto di sopprimere anche informazioni potenzialmente rilevanti.

Fig. 15 - *Estrazione dei prototipi e matching con i features del carattere incognito (Smith et al., 2009)*



Prototipo del carattere

Verde = *Cluster* significativi dai campioni di [h] (per es. per Times Roman)
Blu = *Cluster* confluiti in altri cluster
Magenta = *Cluster* non usati perché coincidono con altri
Rosso = *Cluster* che non contengono suff. campioni per essere significativi

Car. da riconoscere vs Prototipo

Linee tratteggiate = Carattere da riconoscere ([h] eroso)
Linee nere continue = Prototipo
Nero = Buona equivalenza
Magenta = Discreta equivalenza
Ciano = Corrispondenza approssimativa
Rosso = Nessuna corrispondenza

7. *La realizzazione del sistema Navigais con OCR integrato*

La soluzione adottata per il progetto AIS è stato un OCR che fosse: *a*) basato sull'OCR Tesseract; *b*) addestrato specificatamente per il riconoscimento dei simboli dell'Atlante AIS; *c*) integrato con il sistema di navigazione e di interrogazione delle mappe e dei punti di inchiesta di NavigAIS (Fig. 10); *d*) seguito da una fase di post-processing per correggere gli errori eventuali (Fig. 17); *e*) provvisto di una tastiera virtuale (completamente riconfigurabile secondo le esigenze dell'utilizzatore) per l'editing della sequenza restituita da Tesseract e dal post-processing (Fig. 18).

Lo schema a blocchi del sistema realizzato è mostrato in Fig. 10: a sinistra in alto compare la finestra di NavigAIS, che chiameremo Navigatore e che permette di individuare le mappe contenenti un certo argomento nell'indice delle tavole AIS, di caricare la mappa voluta (in basso a sinistra) e di scegliere eventualmente il punto (o i punti) su cui operare il riconoscimento.

Al centro della Fig. 10 figurano i blocchi di elaborazione dell'OCR e di post-elaborazione, ed a destra la finestra dell'*Editor* che serve alla correzione della sequenza fonetica e al salvataggio nel database.

Fig. 16 - *L'OCR opera separando i caratteri dai diacritici. La fase di post-processing si avvale di un set di regole ad hoc, per sopperire alla mancanza di un dizionario ed un modello del linguaggio*

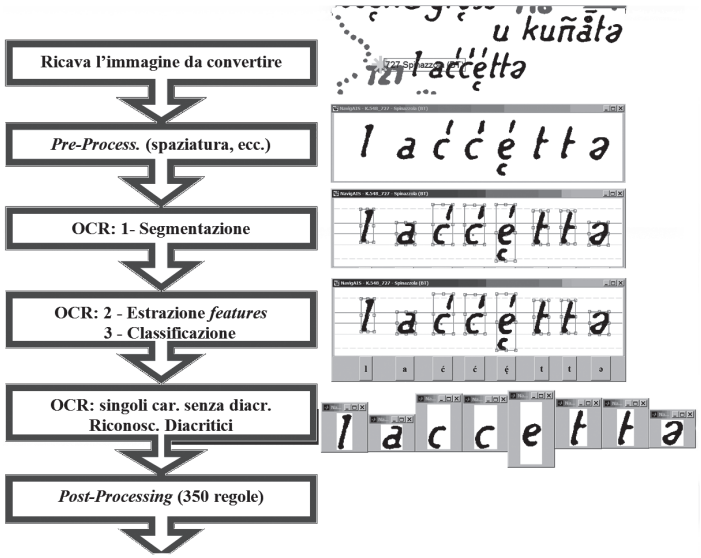
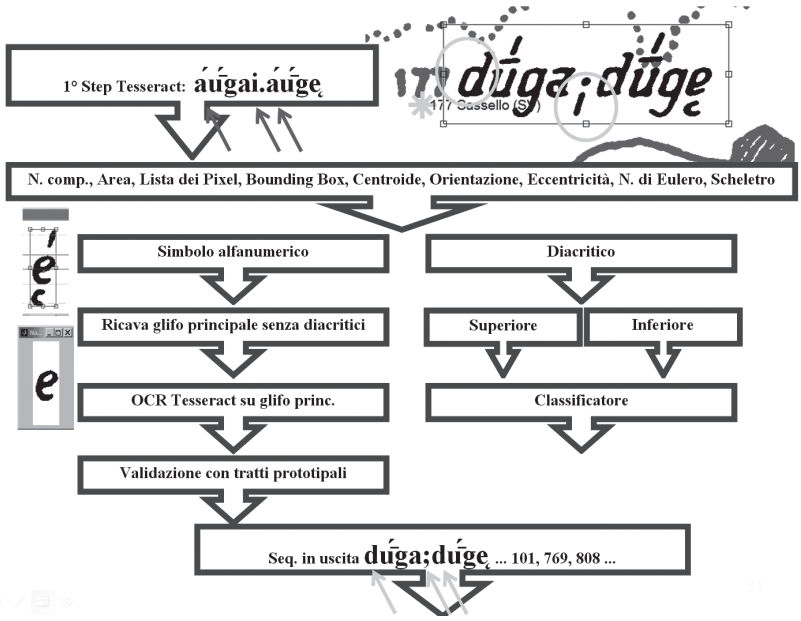


Fig. 17 - *Fase di post-processing. I 12 errori (caratteri + diacritici) commessi dall'OCR Tesseract (in alto a sinistra) sono stati automaticamente corretti dalla fase di post-elaborazione*



Nel processo di acquisizione con l'OCR, il supervisore ha a sua disposizione tre modalità di movimento da un punto all'altro: 1) Il movimento può procedere automaticamente su tutti i punti dell'atlante in sequenza, oppure esclusivamente sui

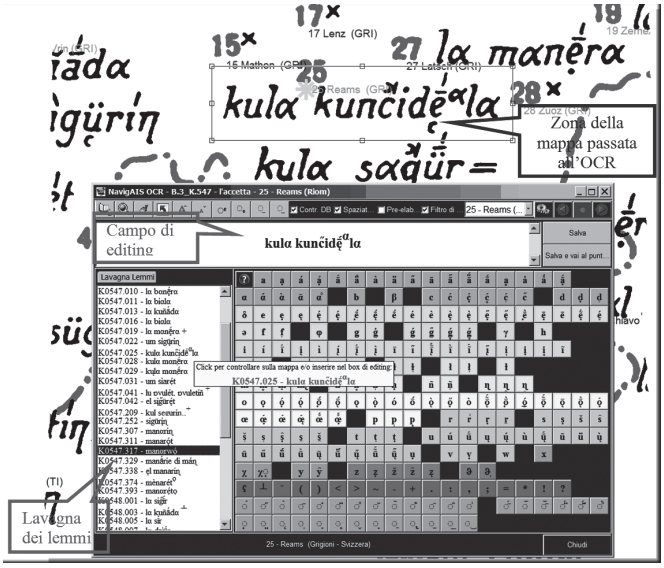
punti appartenenti ad una regione o ad una provincia, oppure infine su una lista di punti selezionati a priori dagli appositi box nell'interfaccia di navigazione, oppure dall'apposito parametro del file di configurazione; 2) il software provvede automaticamente, senza perdita di tempo e senza errori, a spostarsi sui punti di acquisizione voluti secondo un ordine dato; 3) il supervisore svolge due compiti: sceglie quale delle sequenze che circondano il punto sia candidata all'OCR, e controlla nello stesso tempo la correttezza della stringa ritornata dall'OCR.

In alternativa: *a*) l'operatore può selezionare manualmente un punto qualsiasi sulla mappa, per cui un successivo salvataggio avviene nella località che era già selezionata in quel momento; oppure *b*) l'operatore può selezionare un punto cliccando sul nome delle località d'inchiesta, per cui un salvataggio avviene forzatamente all'indirizzo corrispondente.

Dopo la selezione del lemma da acquisire, il programma traccia il riquadro (rettangolo in Fig. 10) con l'area della mappa, in cui i caratteri non siano più distanti di una certa quantità prefissata. Qualsiasi linea tocchi il riquadro esterno è completamente rimossa nel processo di pre-elaborazione, per mantenere solo la sequenza utile al riconoscimento. In questa fase, prima di passare all'elaborazione, l'immagine grafica è salvata su disco, a disposizione per la successiva fase di riconoscimento, ma anche per una eventuale fase di apprendimento dell'OCR.

L'immagine grafica subisce varie fasi opzionali di trattamento, per ripulirla automaticamente dalle porzioni marginali dei glifi, appartenenti ad altre sequenze di caratteri vicini, e per ripararla da piccoli difetti. Se necessario, l'operatore può anche intervenire manualmente, per operare tagli o per "cucire" i rami dei glifi rovinati, nella finestra "Taglia e Cuci" (Fig. 10).

Fig. 18 - *Finestra di lancio dell'OCR e tastiera virtuale per l'editing delle sequenze fonetiche*



Alla fine del riconoscimento, Tesseract restituisce la sequenza fonetica, assieme ai confini dei rettangoli che delimitano ogni singolo carattere (*Bounding Box*), e che possono servire eventualmente a fasi ulteriori di addestramento dell'OCR.

8. *Post-Processing*

Per migliorare la performance, gli OCR comuni ricorrono ad una fase di post-processing che corregge, quando possibile, gli errori dell'OCR, consultando un dizionario ed un modello del linguaggio. Queste componenti possono contribuire anche in maniera essenziale alla fase di segmentazione di una sequenza di caratteri.

Ognuna delle possibili varianti delle parole in uscita dall'OCR è confrontata con il dizionario ed eliminata, se non trova una validazione (ad es. [monte], che esiste, rispetto a [rnonte], che non esiste). Purtroppo, non esiste un dizionario delle parole dialettali dell'Atlante AIS che sia disponibile in un formato accettabile, per poter attribuire una sequenza di caratteri ad una forma certa della lingua in questione. Nel 1960, fu pubblicato postumo (Jaberg era morto nel 1958 e Jud nel 1952) un volume con l'indice delle parole trascritte in maniera tipizzata (ad esempio, "abbághie" invece di [abbáĝĝə]), che non può servire allo scopo che si era prefissato il progetto NavigAIS.

Non esiste neppure un modello del linguaggio (*Statistical Language Model* - SLM), che dia le possibili frequenze di digrammi, trigrammi, ecc., e che risulterebbe particolarmente utile nel riconoscimento della sequenza di caratteri.

Per abbassare la percentuale di errori, si è adottata una specifica strategia per l'Atlante AIS, che ricorre sostanzialmente a due fasi distinte e successive di riconoscimento con Tesseract (Tab. 1):

- Si demanda alla prima fase di riconoscimento svolta da Tesseract il lavoro grossolano di una prima identificazione del carattere, che deve fornire i parametri dei *Bounding Box* indispensabili per le operazioni successive.
- Ottenuta la prima classificazione del carattere, la fase di post-processing provvede a separare il flusso dell'elaborazione in due percorsi diversi (Fig. 17), in cui il primo deve riprocessare e validare la componente di base del glifo da solo con una nuova sottomissione all'OCR Tesseract, mentre il secondo è specializzato nell'elaborare con algoritmi Matlab il gruppo di tutti i diacritici a sé stanti.
- Ogni carattere di base senza diacritici è assoggettato a questo secondo passaggio solitario attraverso Tesseract, con l'uscita ristretta unicamente ai segni alfanumerici senza diacritici, per ridurre contemporaneamente sia il numero dei glifi a soli 91 simboli, sia le possibili ambiguità fra caratteri, e di conseguenza il CER complessivo.
- I diacritici sono separati dal corpo principale del glifo, sono divisi fra diacritici in apice e a pedice, e classificati con regole basate sulle loro caratteristiche topologiche, sulla loro compatibilità reciproca ed anche sulla loro compatibilità con il carattere di base. In questo modo la complessità della casistica infinita delle forme (nel caso delle vocali circa 8480) e dei livelli dei diacritici, si riduce a pochi casi semplici (simboli alfanumerici + diacritici) in numero inferiore a 150.

Ovviamente gli errori sono ancora possibili per i glifi scritti male, rovinati, fusi con i diacritici, e anche per forme male addestrate, ecc.

– Infine, l'ultimo stadio della post-elaborazione applica circa 350 regole di controllo dell'identità dei caratteri, basate sulla topologia, le dimensioni e la conformità fonetica delle componenti dei glifi. In questa fase, per aumentare la velocità di elaborazione, i simboli sono classificati ed elaborati in 5 categorie separate: 1 – Vocali, 2 – Consonanti, 3 – Cifre, 4 – Punteggiatura, 5 – Segni diacritici. Se ad esempio l'OCR restituisce una [i], ma il numero delle componenti reali, ricavato dai pixel dell'immagine, è invece di una sola componente, il carattere, a seconda delle dimensioni e della posizione del baricentro, può essere una [ɪ] oppure una [,]. Se invece le componenti sono due, ma il baricentro è basso (sotto la media risultante dei caratteri AIS) allora si tratta di un [;]. Altrettanto dicasi ad es. di una [é] accentata, che abbia un'unica componente, con la conseguente eliminazione dell'accento.
  Se l'OCR ritorna un [.], ma la componente ha un buco (numero di Eulero = 0), allora (a seconda delle dimensioni e della posizione) può trattarsi di un diacritico [o̊], oppure [o̩], oppure del carattere [o], oppure di [o] in apice [ᵒ], ecc. Un altro tipo di controllo interessa anche l'associazione dei diacritici e dei glifi di base, che risulti accettabile per l'inventario fonetico dell'AIS.

Si deve notare che queste regole non sono per nulla definitive e sono state, e possono essere ancora, modificate, raffinate, ed estese in continuazione, sulla base dell'esperienza e di altre conoscenze rilevanti acquisite nel corso del processo di riconoscimento.

Tab. 1 - *Schema delle regole applicate nella fase di post-processing*

Tipologia	**Scopo dell'elaborazione**
N. componenti	Per permettere, ad es., di discriminare: [i] da [ɪ], [;] da [,].
Area	Per distinguere fra oggetti irrilevanti, diacritici e caratteri.
Lista dei Pixel	Per individuare le componenti connesse.
Bounding Box	Per confronti differenziali basati sulle dimensioni.
Estremi	Per distinguere, ad es., le parentesi.
Centroide	Per la dilatazione dei caratteri, linee di riferimento.
Orientazione	Per distinguere, ad es., accento primario [ó] da secondario [ò].
Eccentricità	Per discriminare componenti lineari (linea=1) da ellittici (cerchio=0): ad es. [ō], [o̅], rispetto a [o̭].
N. di Eulero	Per distinzione basata sul numero di "fori" del componente: ad es. [e] da [œ], [g] da [y], [o̩] da [o̊], ecc.
Angoli	Per discriminare alcuni diacritici.
Scheletro	Per distinguere diacritici e caratteri: ad es. [u] da [y], [k] da [h].

9. *Valutazione dei risultati dell'OCR*

Per la valutazione della prestazione del riconoscitore, sono stati realizzati vari test che hanno riguardato sia le tappe del *training* dell'OCR, sia la fase di riconosci-

mento vero e proprio. Per la fase di addestramento, è stato preparato un test minimale con 50 lemmi (circa 500 caratteri complessivamente) provenienti dal database AMDV.

Il test è stato volutamente limitato, per testare rapidamente l'effetto delle modifiche che si sono succedute nel tempo, e per confrontare fra di loro l'accuratezza dei vari linguaggi addestrati e di tutti i linguaggi fra di loro. La Fig. 19 riassume i risultati dell'OCR con linguaggi addestrati su diversi set di *training*, presi a sé stanti oppure in combinazione fra di loro. Le tre colonne verticali riportano il CER (*Character Error Rate*) per le tre modalità di test principali (Fig. 19-20): 1) OCR Tesseract da solo senza dilatazione dei caratteri (a destra); 2) OCR Tesseract da solo con dilatazione dei caratteri (al centro); 3) OCR con post-processing (a sinistra).

Fig. 19 - *Risultati comparativi di vari test fatti con i linguaggi addestrati su diversi set di training. Le tre colonne riportano il CER per le tre modalità di test principali: OCR Tesseract da solo senza dilatazione dei caratteri (a destra), OCR Tesseract da solo con dilatazione dei caratteri (al centro), OCR con post-processing (a sinistra)*

	NavigAIS OCR Test											
	Char Dilatation + Post_Processing				Char Dilatation				OCR only			
	OCR Language	CER	WER	WER (	OCR Language	CER	WER	WER (	OCR Language	CER	WER	WER (
5	ais01+ais12+ais09+ais11_dilat_post	1,41	6,6	6,6	ais01+ais12+ais09+ais11_dilat	22,05	73,58	70,75	ais01+ais12+ais09+ais11_no_dil	24,51	71,7	66,98
6	ais01+ais12+ais09_dilat_post	1,41	6,6	6,6	ais01+ais12+ais09_dilat	22,05	73,58	69,81	ais01+ais12+ais09_no_dil	25,04	74,53	68,87
7	ais01+ais12+ais10+ais11_dilat_post	1,41	6,6	6,6	ais01+ais12+ais10+ais11_dilat	22,57	84,91	81,13	ais01+ais12+ais10+ais11_no_dil	26,46	79,25	73,58
8	ais01+ais12+ais10_dilat_post	1,41	6,6	6,6	ais01+ais12+ais10_dilat	22,93	83,96	81,13	ais01+ais12+ais10_no_dil	26,63	80,19	74,53
9	ais01+ais09+ais11+ais12_dilat_post	1,41	6,6	6,6	ais01+ais09+ais11+ais12_dilat	23,1	76,42	72,64	ais01+ais09+ais11+ais12_no_dil	25,4	74,53	69,81
10	ais01+ais09+ais12+ais11_dilat_post	1,41	6,6	6,6	ais01+ais09+ais12+ais11_dilat	23,1	76,42	72,64	ais01+ais09+ais12+ais11_no_dil	25,4	74,53	69,81
11	ais12+ais01+ais10+ais11_dilat_post	1,41	6,6	6,6	ais12+ais01+ais10+ais11_dilat	23,1	88,68	83,96	ais12+ais01+ais10+ais11_no_dil	27,16	83,02	77,36
12	ais12+ais01+ais09+ais11_dilat_post	1,41	6,6	6,6	ais12+ais01+ais09+ais11_dilat	23,81	81,13	77,36	ais12+ais01+ais09+ais11_no_dil	26,81	76,42	71,7
13	ais01+ais12+ais09+ais10_dilat_post	1,41	6,6	6,6	ais01+ais12+ais09+ais10_dilat	24,34	76,42	74,53	ais01+ais12+ais09+ais10_no_dil	26,63	75,47	70,75
14	ais01+ais12+ais10+ais09_dilat_post	1,41	6,6	6,6	ais01+ais12+ais10+ais09_dilat	24,69	78,3	76,42	ais01+ais12+ais10+ais09_no_dil	26,98	76,42	72,64
138	ais09+ais10+ais11+ais01_dilat_post	2,12	10,38	9,43	ais09+ais10+ais11+ais01_dilat	31,75	85,85	81,13	ais09+ais10+ais11+ais01_no_dil	32,45	86,79	82,08
139	ais01+ais12_dilat_post	2,47	9,43	7,55	ais01+ais12_dilat	19,58	77,36	74,53	ais01+ais12_no_dil	22,22	75,47	71,7
140	ais11+ais10_dilat_post	2,47	12,26	12,26	ais11+ais10_dilat	29,63	83,02	80,19	ais11+ais10_no_dil	29,1	83,96	78,3
141	ais01+ais12+ais11_dilat_post	2,65	10,38	7,55	ais01+ais12+ais11_dilat	19,22	76,42	75,47	ais01+ais12+ais11_no_dil	21,52	72,64	69,81
142	ais12+ais11_dilat_post	2,65	12,26	9,43	ais12+ais11_dilat	21,52	88,68	85,85	ais12+ais11_no_dil	23,99	84,91	83,02
143	ais11+ais09_dilat_post	2,82	13,21	13,21	ais11+ais09_dilat	28,22	83,02	79,25	ais11+ais09_no_dil	29,1	83,02	78,3
144	ais10+ais09_dilat_post	2,82	14,15	13,21	ais10+ais09_dilat	32,63	86,79	82,08	ais10+ais09_no_dil	32,98	85,85	82,08
145	ais12_dilat_post	3,17	15,09	13,21	ais12_dilat	24,69	99,06	96,23	ais12_no_dil	25,4	92,45	88,68
146	ais10_dilat_post	3,17	16,04	16,04	ais10_dilat	31,22	89,62	83,96	ais10_no_dil	31,57	90,57	84,91
147	ais09_dilat_post	3,88	18,87	17,92	ais09_dilat	34,04	92,45	86,79	ais09_no_dil	34,39	89,62	85,85
148	ais11_dilat_post	6,7	29,25	26,42	ais11_dilat	27,87	85,85	82,08	ais11_no_dil	29,63	85,85	83,02
149	ais01+ais11_dilat_post	7,23	24,53	21,7	ais01+ais11_dilat	20,63	63,21	61,32	ais01+ais11_no_dil	22,57	67,92	64,15
150	ais01_dilat_post	11,82	44,34	39,62	ais01_dilat	21,52	68,87	63,21	ais01_no_dil	24,69	74,53	67,92
151	ais_f_dilat_post	25,93	93,4	89,62	ais_f_dilat	40,56	116,04	109,43	ais_f_no_dil	40,56	108,49	102,83
153	Media errore con post-processing	1,735643			Media con dilat. caratteri	26,47582			Media solo OCR	28,12795		

Il risultato migliore (errore dell'1.41%) è ottenuto dalla combinazione di vari linguaggi (fra cui ais01, ais09, ais10, ais11, ais12) e dall'esecuzione del post-processing, mentre il peggiore, CER = 25.93%, è dato dal linguaggio allenato sul set costruito artificialmente con caratteri di un font che doveva imitare l'AIS, ma che evidentemente è abbastanza lontano dalla realtà grafica delle tavole AIS.

Senza la fase di post-processing, ma con la dilatazione dei caratteri, il CER aumenta e varia dal 22% al 40%, mentre senza la fase di post-processing, e senza la dilatazione dei caratteri, il CER aumenta dal 24% al 40%, con un peggioramento del 2-3% rispetto al caso della dilatazione dei caratteri. La Fig. 19 riporta anche il WER (*Word Error Rate*), l'errore sulle parole intere, che non ha molto interesse per l'AIS.

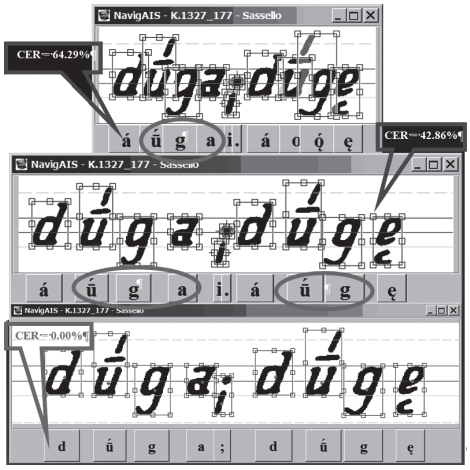
Il tempo di acquisizione per punto è risultato inferiore ai 14" per sequenze di circa 8.36 caratteri, più che dimezzato rispetto ai 30", ipotizzati inizialmente nel

progetto. Si deve ovviamente tenere conto della lunghezza della sequenza, per cui all'aumentare del numero di caratteri il tempo di acquisizione aumenta di qualche secondo nelle frasi più lunghe.

Per quanto riguarda il tasso di errore, invece, questo non dipende dalla lunghezza della sequenza riconosciuta dall'OCR e rimane invariato. L'errore può aumentare di qualche punto percentuale nelle sequenze "anomale", in cui i caratteri siano appiccicati e nel caso il deterioramento dei caratteri sia alto. In questi casi, in effetti, la mancanza di un dizionario dei lemmi dialettali impedisce all'OCR di interpretare correttamente i caratteri lesionati in una certa sequenza.

Un test di acquisizione su dati reali provenienti da 14 tavole complete per 100.000 caratteri complessivi ha dato un CER medio del 3.65%, includendo nel calcolo una qualsiasi inserzione, omissione e sostituzione dei caratteri, diacritici compresi (con la cosiddetta *Distanza di Levenshtein*; Levenshtein, 1965).

Fig. 20 - *Risultati OCR: Senza distanziamento dei caratteri, l'errore CER commesso da Tesseract è del 64.29%. Con il distanziamento, ma senza post-processing sul risultato dell'OCR, l'errore CER si riduce al 42.86%. Infine con distanziamento e post-elaborazione, tutti i caratteri sono riconosciuti correttamente con CER = 0% (NavigAIS-web K. 1327.177)*



10. *Conclusioni*

Basandosi sulla percentuale di errori CER (*Character Error Rate*), ottenuta nei test e sui dati reali di 14 mappe per circa 100.000 caratteri (CER medio inferiore al 3.65%, molto minore di quello ipotizzato inizialmente), e sui tempi di acquisizione per punto (inferiore ai 14" per sequenze con una media di 8.36 simboli, dimezzato rispetto alle stime iniziali), si può ragionevolmente predire che il completamento dell'acquisizione dell'intero AIS avverrà in meno di 2 anni/uomo. Queste stime sono confermate dai risultati già ottenuti, per cui alla fine del 2019 il 50% delle mappe AIS è già stato completamente trascritto.

Bibliografia

Goebl, H. (1994). *L'Atlas linguistique du ladin central et des dialectes limitrophes (première partie, ALD-I)*. In Mouton, P.G. (Ed.), *Geolingüística. Trabajos europeos*. Madrid: Consejo Superior de Investigaciones Científicas, 155-168.

Jaberg, K., Jud, J. (1928). *Der Sprachatlas als Forschungsinstrument. Kritische Grundlegung und Einführung in den Sprach- und Sachatlas Italiens und der Südschweitz*. Halle: Max Niemeyer. [Traduzione italiana a cura di G. Sanga: *Ais - Atlante Linguistico ed Etnografico dell'Italia e della Svizzera Meridionale. Vol. 1: Fondamenti Critici e Introduzione, Vol. 2: Scelta di Carte Commentate*. Milano: Unicopli, 1988.]

Jaberg, K., Jud, J. (1928-1940). *Sprach- und Sachatlas Italiens und der Südschweiz, Vol. 1-8*. Ringier, Zofingen, Bern: Max Niemeyer. [Ristampa 1971-1981: Nendeln/New York: Kraus Reprint]. www3.pd.istc.cnr.it/navigais-web.

Jaberg, K. (1960). *Index zum Sprach- und Sachatlas Italiens und der Südschweiz: Ein propädeutisches etymologisches Wöterbuch der italienischen Mundarten.* Berna: Stämpfli.

Kattenbusch, D. (1995). *Atlas parlant de l'Italie par régions: VIVALDI*. In Aa. Vv. *Estudis de lingüística i filologia oferts a Antoni M. Badia i Margarit*. Barcellona: Universitat Autònoma de Barcelona, Departamento de Filología Catalana y de Filología Española, 443-455. www.dailybest.it/web/vivaldi-mappa-interattiva-dialetti-italiani.

Kramer, J. (1988-1998). *Etymologisches Wörterbuch des Dolomitenladinischen (EWD), voll. 1-8*. Amburgo: H. Buske Verlag.

Loporcaro, M. (2014). *AIS Reloaded*, https://www.rose.uzh.ch/de/forschung/forschungamrose/projekte/AIS-reloaded.html.

Mott, A., Kezich, G. & Tisato, G. (2003). *Il Trentino dei contadini. Piccolo atlante sonoro della cultura materiale.* San Michele all'Adige (TN): Museo degli Usi e Costumi della Gente Trentina. www.museosanmichele.it/editoria/shop/opere-varie-volumi-esauriti/g-kezich-a-mott-g-tisato-il-trentino-dei-contadini-cdrom.

Scheuermeier, P. (1943). *Bauernwerk in Italien, der italienischen und rätoromanischen Schweiz, Vol. 1-2*. Erlenbach/Zurigo: Rentsch. [Traduzione italiana: Scheuermeier P. (1980). *Il lavoro dei contadini*. Milano: Longanesi.]

Telmon, T., Canobbio, S. (1985). *Atlante linguistico ed Etnolinguistico del Piemonte Occidentale.* Torino: Regione Piemonte.

Tisato, G. (2010). NavigAIS – AIS Digital Atlas and Navigation Software. In *Atti del VI Convegno AISV*, Napoli, 3-5 febbraio 2010, 451-461 (versione scaricabile: www3.pd.istc.cnr.it/navigais, versione online: www3.pd.istc.cnr.it/navigais-web).

Tisato, G., Vigolo, M.T. (2011). *Atlante Multimediale dei Dialetti Veneti (AMDV)*. In Borgato, G., Vanelli, L. (Eds.), *Atti del Convegno "In ricordo di Alberto Zamboni"*. Padova, 25, gennaio, 2011, 99-126.

Tisato, G., Barbierato, P., Ferrieri, G., Gentili, C. & Vigolo, M.T. (2013). Atlante Multimediale dei Dialetti Veneti. In *Atti del IX Convegno AISV 2013, Multimodalità e Multilingualità*, Università Ca' Foscari, Venezia, 21-21 gennaio 2013. Roma: Bulzoni, 445-462.

Tisato, G., Vigolo, M.T. (2016). Dagli Atlanti storici agli Atlanti multimediali: il NavigAIS e l'AMDV (Atlante Multimediale dei Dialetti Veneti). In *Atti del Convegno*

Internazionale di Studi Archivi Etnolinguistici Multimediali, Museo delle Genti d'Abruzzo, Pescara, 6 ottobre 2012, 96-123.

Vithlani, P., Kumharana, C.K. (2015). Structural and Statistical Feature Extraction Methods for Character and Digit Recognition. In *International Journal of Computer Applications*, 120, 43-47.

Zamboni, A. (1974). *I dialetti veneti*. Pisa: Pacini.

Zamboni, A. (1984). I dialetti cadorini. In Pellegrini, G.B., Sacco, S. (Eds.), *Atti del Convegno Internazionale "Il ladino bellunese"*. Belluno, 2-4 giugno 1983, 45-83.

Zamboni, A., Vigolo, M.T. (2011). Tra nomi e cose. Commenti lessicali e onomasiologici allo Scheuermeier veneto. In Perco, D., Sanga, G. & Vigolo, M.T. (Eds.), Scheuermeier, P., *Il Veneto dei contadini 1921-1932*. Vicenza: Angelo Colla Editore, 67-87.

Zipf, G.K. (1949). *Human Behaviour and the Principle of Least-Effort*. Oxford: Addison-Wesley Press.

Ringraziamenti

Un ringraziamento a tutto il team del progetto *AIS Reloaded*, in particolare a Chiara Zanini, Giulia Donzelli e Stefano Negrinelli, che stanno lavorando alla raccolta dei dati nella Svizzera Romanza ed hanno collaborato (assieme a Giacomo Ferrieri) alla preparazione dell'inventario fonetico AIS.