

Acquisizione Digitale dell'Intero AIS

GRAZIANO TISATO

graziano@tisato.it

Istituto di Scienze e Tecnologie della Cognizione (ISTC)

Consiglio Nazionale delle Ricerche (CNR)

Via Martiri della Libertà, 2 – 35127 Padova

Si presenta in questo articolo la terza fase del progetto NavigAIS, nato nel 2009, con lo scopo di acquisire digitalmente il testo dell'intero AIS (www3.pd.istc.cnr.it/navigais, www3.pd.istc.cnr.it/navigais-web), l'Atlante Italo-Svizzero dei linguisti svizzeri K. Jaberg e J. Jud (*Sprach- und Sachatlas Italiens und der Südschweiz*, 1928-1940), e di mettere a disposizione della comunità il relativo database.

Il lavoro si è svolto nell'ambito del progetto AIS Reloaded (Romanisches Seminar, Università di Zurigo, www.rose.uzh.ch/de/forschung/AIS-reloaded.html), finanziato da SNSF (Swiss National Science Foundation), che oltre alla digitalizzazione dell'AIS intende analizzare l'evoluzione diacronica dei dialetti italo-romanzi avvenuta in cent'anni nel Canton Ticino e nei Grigioni, con una campagna di registrazioni nelle medesime località, oggetto dell'inchiesta AIS, e sullo stesso questionario (intermedio) usato a quel tempo.

L'articolo affronta le problematiche del simbolismo AIS, l'architettura e le funzionalità dell'OCR (Optical Character Recognition) realizzato *ad hoc* fra il 2016 e il 2017, la metodologia usata per la validazione dei risultati ed il controllo dei lemmi del database, la valutazione obiettiva delle performance del riconoscitore (considerando tutti i caratteri errati, omissi e spuri), ed i tempi di acquisizione dei lemmi.

Si presenta anche, come primo risultato del lavoro, l'inventario fonetico effettivo dell'AIS ricavato dai dati di circa il 25% delle mappe AIS, funzionale per il miglioramento delle performance dell'OCR, in quanto consente di individuare sia i caratteri errati che le combinazioni di caratteri e diacritici non permessi nella sequenza restituita dall'OCR, e di accorciare i tempi di acquisizione del 10% circa.

The screenshot displays the NavigAIS application interface. At the top, a table lists AIS maps (K0545 to K0548) with columns for map number, Italian name, German name, French name, legend, and comment. Below the table, a map of Switzerland shows AIS points (e.g., 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100) and their corresponding AIS points. A virtual keyboard is visible at the bottom left, and a search bar is at the bottom right. Red arrows point from text labels to specific features: 'Database dei lemmi salvati' points to the table; 'Contr. del lemma sulla mappa' points to the map; 'Parametri e opzioni OCR' points to the keyboard; 'Lavagna temporanea dei lemmi acquisiti' points to the search bar; 'Tastiera virtuale configurabile' points to the keyboard.

1	Num. Mappa AIS	K0545	K0546	K0547	K0548	K0548
2	Nome in Italiano	fascina	io della f	accetta	scura	da squa
3	Deutsch Name			heil	axl	
4	Nom Français			hache	cognée	
5	Legende (De)			Karte0547/Legende.doc	Karte0548/Legende.doc	
6	Legenda (It)			Karte0547/K0547.htm	Karte0548/K0548.htm	
7	Commento					
8	N. Luogo Inchiesta AIS Punto AIS					
9	1	Brigels-Brell	1		la biala	la sigir
10	2	Pitasch	3		la biala	la kuhada⁺
11	3	Fino Domest	5		la biala	la sir

Fig. 1 – NavigAIS: versione con OCR. Sullo sfondo, la tavola AIS 548 ([NavigAIS-web 548.71](http://NavigAIS-web.548.71)).

A sinistra, in basso, la tastiera virtuale (configurabile) per la correzione degli eventuali errori del riconoscitore.

In alto, il database in cui le sequenze riconosciute sono automaticamente salvate nella locazione opportuna.

Il progetto NavigAIS si è sviluppato in tre fasi distinte:

- La prima fase è stata completata nel 2009 con la realizzazione della versione digitale e navigabile dell'AIS (scaricabile a www3.pd.istc.cnr.it/navigais), resasi indispensabile per portare a termine l'Atlante Multimediale dei Dialetti Veneti (Tisato et al., 2013, www.pd.istc.cnr.it/amdv). Le mappe sono state preliminarmente separate su due livelli sovrapponibili, l'uno contenente il solo testo, e l'altro i confini geografici e l'orografia, in maniera tale che non interferissero poi con il riconoscimento da parte dell'OCR.

- La seconda fase (2014) ha riguardato l'implementazione della versione *online* dell'AIS (www3.pd.istc.cnr.it/navigais-web), per permetterne la consultazione diretta *web*, e l'accesso facilitato ad una mappa e ad un punto AIS, dando un opportuno indirizzo in una qualsiasi pagina HTML (ad es. [mappa 1327 punto 177](#)).
- La terza fase è stata portata a termine fra 2016 e 2017, con la creazione di un opportuno OCR (Fig. 1-2), per permettere la digitalizzazione del testo dell'intero atlante AIS (metà delle 1705 mappe saranno completate entro la fine del 2019).

In generale, le sequenze fonetiche dell'AIS, che devono trascrivere lingue di regioni diverse, sono caratterizzate da un numero molto alto di caratteri a bassa frequenza, con un andamento che può essere descritto dalla legge di Zipf (*Human Behaviour and the Principle of Least-Effort*, 1949). Conseguenza inevitabile di questo andamento è il peggioramento dei risultati di un qualsiasi OCR basato su un motore probabilistico.

Il numero di simboli diversi da riconoscere (circa 1400), la scrittura manuale, il corsivo (circa 70° di inclinazione) dei caratteri, la struttura dei diacritici, articolata su 7 livelli, e l'incertezza di assegnazione di un lemma ad un punto AIS rendono inutilizzabili gli OCR (si veda [S. Lücke, 2018](#)).

Manca d'altra parte per le trascrizioni AIS un modello del linguaggio, che possa predire le probabilità di successione di un carattere (e bigrammi, trigrammi, ecc.), così come manca un dizionario delle parole esistenti nelle lingue in questione, che permetta di individuare la parola corretta fra le possibili sequenze fornite dall'OCR.

L'approccio innovativo adottato per il progetto AIS è stato la riduzione della complessità del riconoscimento, con un'elaborazione che separa automaticamente il glifo di base dei caratteri dai relativi diacritici sovrastanti e sottostanti, per cui i simboli distinti da riconoscere calano a circa 70 simboli vocalici e 134 consonantici.

Il glifo di base può essere in questo modo trattato con un riconoscitore tradizionale (in questo caso è stato scelto Tesseract, perché è di libero uso e perché mette a disposizione il codice).

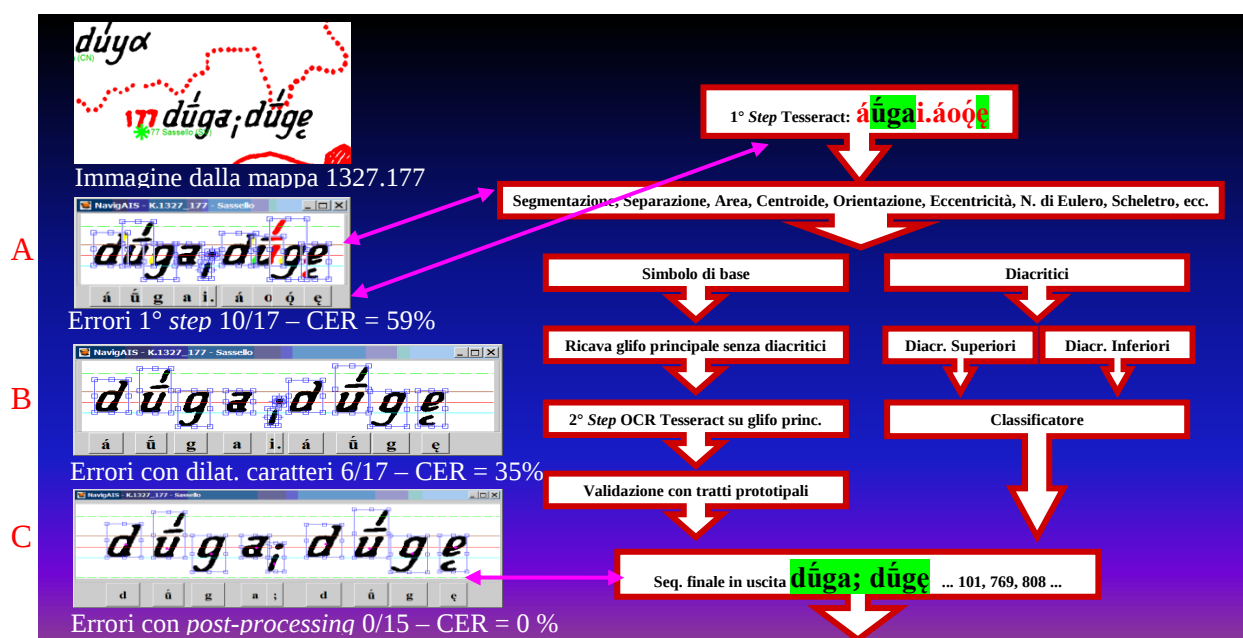


Fig. 2 – Sulla destra lo schema dell'OCR implementato. Sulla sinistra un es. di errori (contando sia i glifi di base che i diacritici restituiti dall'OCR) in uscita dal 1° step da Tesseract (A), e il miglioramento che avviene con la dilatazione dei caratteri (B), e con le correzioni automatiche nelle fasi finali di *post-processing* (C).

Tesseract ha subito un addestramento preliminare su un campione limitato di frasi etichettate. A quel punto inizia il processo di riconoscimento che automaticamente percorre i 407 punti dell'inchiesta AIS per ognuna delle 1705 mappe dell'atlante. Il primo passaggio serve principalmente a segmentare la catena fonetica nei singoli caratteri e relativi diacritici (Fig. 2A). Segue una fase di *pre-processing* con un software Matlab, per spaziare opportunamente i caratteri e migliorare il tasso di riconoscimento (Fig. 2B), e per l'estrazione di varie caratteristiche geometriche e topologiche, e di misure utili ad alimentare lo step intermedio, ed infine il processo si completa con la fase finale di *post-processing* (Fig. 2C).

Nella fase centrale i glifi senza diacritici sono assoggettati nuovamente all'OCR Tesseract, mentre i diacritici seguono un percorso apposito e sono classificati in base alle proprietà geometriche e topologiche elementari.

Nel passo finale, caratteri e diacritici sono validati e corretti con una serie di 350 regole implementate con Matlab, ed infine la sequenza è salvata in un *database*. In quest'ultima fase, un supervisore può intervenire per correggere eventuali errori con la tastiera (configurabile dall'utente in base alle esigenze) di Fig. 1.

Per quanto riguarda i risultati (che saranno esposti in dettaglio nell'articolo), si riassume qui solo la percentuale di errore CER (*Character Error Rate*), che si è ottenuta su un campione rappresentativo di 14 mappe AIS complete per circa 6000 parole complessive (100000 caratteri), e che varia da 1.35% a 5.87% per mappa con valore medio del 3.65% (molto inferiore a quanto inizialmente ipotizzabile per questa tipologia di caratteri), e con tempi di acquisizione (ed eventuali correzioni) inferiori alle 2.30 ore per mappa, il che dà sufficienti garanzie che il completamento dell'acquisizione dell'intero AIS possa avvenire in circa due anni/uomo.