

ALMA MATER STUDIORUM – UNIVERSITÀ DI BOLOGNA

**DIPARTIMENTO DI INTERPRETAZIONE E
TRADUZIONE**

CORSO di LAUREA IN
MEDIAZIONE LINGUISTICA INTERCULTURALE (Classe L-12)

ELABORATO FINALE

**Verso lo sviluppo di un modello predittivo per lo
screening del Disturbo di Linguaggio in età
evolutiva: un esperimento-pilota con Orange**

CANDIDATO:

Francesco Fernicola

RELATORE:

Mazzoleni Marco

ANNO ACCADEMICO: 2018-2019

Secondo appello

Introduzione

Sin dai suoi albori, la linguistica computazionale si è sviluppata come un campo fondamentalmente interdisciplinare, sfruttando gli innovativi strumenti informatici sorti durante la rivoluzione digitale per analizzare le lingue naturali e i fenomeni linguistici.

Questo campo ha fornito un contributo fondamentale per la creazione di alcuni tra gli strumenti che sono diventati parte integrante della vita di tutti i giorni; a partire dagli editor di testo (Microsoft Word), passando per i motori di ricerca (Google) fino ai sistemi di riconoscimento vocale e di traduzione automatica. Questi ultimi due in particolare hanno conosciuto miglioramenti notevoli negli ultimi anni, grazie agli enormi progressi nelle tecniche di *machine learning*, anche noto come *apprendimento automatico*.

Tuttavia queste tecniche non trovano la propria applicazione solo nella nostra quotidianità, ma anche nella ricerca scientifica. È infatti possibile utilizzarle per l'analisi di dati e persino per la creazione di modelli predittivi a partire da una collezione di dati (chiamata *dataset*). Rappresentano pertanto uno strumento aggiuntivo per la valutazione delle ipotesi scientifiche, nonché un supporto ulteriore alla diagnosi in campo medico.

In questa tesi si intende fornire un'applicazione pratica di queste tecniche al campo della logopedia, proponendo un esperimento-pilota con il fine di sviluppare un modello predittivo per lo screening del *Disturbo del Linguaggio* in età evolutiva.

Nel primo capitolo di questa tesi verranno introdotte le nozioni fondamentali per poter comprendere cosa viene inteso come *Disturbo del Linguaggio*, esaminando le caratteristiche che lo individuano e le sfide nella sua definizione.

Verranno successivamente descritti nel secondo capitolo gli strumenti informa-

tici utilizzati ai fini della raccolta, preparazione ed elaborazione dei dati linguistici, grazie anche al supporto delle schermate dei software.

Nell'ultimo capitolo verrà discusso lo studio condotto dalla AUSL Toscana, che ha fornito l'ispirazione per questa tesi, e verranno descritti i vari passaggi dell'esperimento in oggetto.

Indice

Introduzione	i
1 Inquadramento Teorico	1
1.1 Definire il Disturbo del Linguaggio	1
1.1.1 Criteri di esclusione	2
1.1.2 Criteri di inclusione	3
1.2 Aree linguistiche	4
1.2.1 Fonologia	4
1.2.2 Morfosintassi	5
1.2.3 Sintassi	5
1.2.4 Semantica lessicale e frasale	5
1.2.5 Pragmatica	6
1.3 Perché intervenire tempestivamente	6
2 Strumenti utilizzati	9
2.1 ELAN	9
2.2 Orange	10
3 Metodologia	13
3.1 Studio clinico sulle abilità narrative di bambini affetti da DSL	13
3.1.1 Metodo	14
3.1.2 Risultati	15
3.2 Preparazione dei dati	16
3.3 Esperimento con Orange	16
3.3.1 L'algoritmo "kNN"	17

3.3.2	Classificazione	17
3.3.3	Regressione	19
3.3.4	Risultati	21
	Conclusioni	23
	Bibliografia	25

Elenco delle figure

2.1	Interfaccia di ELAN	10
2.2	Esempio di un <i>workflow</i> con il <i>widget</i> "File"	11
3.1	Interfaccia <i>widget</i> "Test & Score"	18
3.2	<i>Workflow</i> per il test di classificazione	19
3.3	<i>Workflow</i> per il test di regressione	20
3.4	Predizioni effettuate con kNN	20

Capitolo 1

Inquadramento Teorico

Al fine di sviluppare un algoritmo di *machine learning* che sia in grado di svolgere delle predizioni affidabili è innanzitutto necessario inquadrare la questione in maniera adeguata da un punto di vista teorico.

In questo capitolo verrà innanzitutto data la definizione di *Disturbo Primario del Linguaggio (DSL)* seguendo le linee guida sviluppate durante la Consensus Conference promossa dall'*Associazione CLASTA* e dalla *Federazione Logopedisti Italiani* (M Chiara Levorato e Rossetto 2018). Successivamente verrà fornita un'analisi nel dettaglio dei criteri di esclusione e inclusione, per poi passare a una breve descrizione delle singole aree linguistiche compromesse dai DSL.

Infine, verranno fornite le motivazioni per la necessità di intervenire in anticipo e le nuove possibilità offerte in questo ambito dagli ultimi sviluppi nel campo dell'informatica.

1.1 Definire il Disturbo del Linguaggio

Con il termine "Disturbo Primario del Linguaggio" si definisce una gamma di condizioni perlopiù eterogenee che rappresentano i disordini dello sviluppo più frequenti in età evolutiva. Essi possono infatti essere riscontrati in diverse aree del linguaggio, limitandosi in alcuni casi alla produzione, mentre in altri estendendosi anche alle aree della comprensione ed elaborazione linguistica, a partire dalla base fonetico-

articolatoria fino ad arrivare talvolta anche a quella pragmatico-discorsiva (Guasti e Marotta 2018).

Da un punto di vista terminologico, si evidenzia che all'espressione "Disturbo Primario del Linguaggio" viene associata l'abbreviazione "DSL" in quanto per anni è stata in uso l'espressione "Disturbo Specifico del Linguaggio". Secondo gli studiosi, con questa definizione si è tuttavia in passato commesso l'errore di ritenere il disturbo limitato alla sola area linguistica, ignorando le frequenti osservazioni cliniche di comorbidità con altri deficit in aree come la coordinazione motoria o l'attenzione (ADHD) (Bishop et al. 2017).

Pertanto è emersa l'intenzione di evitare di sottolinearne la specificità, preferendo il termine "Disturbo Primario del Linguaggio", il quale non è tuttavia ancora stato adottato in maniera del tutto uniforme. Ciò ha portato le due espressioni a essere spesso intercambiabili in letteratura, venendo infatti utilizzate in maniera pressoché indiscriminata nel testo della Consensus Conference (Guasti e Marotta 2018).

Al fine di effettuare una corretta diagnosi di disturbo primario del linguaggio è prima di tutto necessario delineare esattamente quelli che vengono definiti come "criteri di esclusione" e "criteri di inclusione".

1.1.1 Criteri di esclusione

L'età di 4 anni è innanzitutto definita come determinante ai fini della diagnosi, in quanto prima di allora non è possibile diagnosticare un disturbo con assoluta certezza (ibid.). Successivamente è necessario escludere tutte le altre possibili ragioni per cui tale ritardo nella produzione possa presentarsi. Esistono infatti numerose altre condizioni che possono far insorgere problemi nell'area linguistica, pur non essendo strettamente legate a essa.

Non vengono pertanto considerati affetti da DSL "[...] i bambini con sordità, con deficit neurologici acquisiti, con disturbi dello spettro autistico, con ADHD [...]"(ibid.). Non solo, è anche necessario che presentino un QI non verbale entro i parametri della loro età (<85 punti) in congiunzione con problemi in almeno una

delle svariate aree linguistiche (fonologia, morfosintassi, sintassi, semantica lessicale e frasale, pragmatica) (ibid.).

1.1.2 Criteri di inclusione

Mentre i criteri di esclusione sono ben definiti e solitamente condivisi dalla maggior parte della comunità scientifica, l'opinione riguardo ai criteri di inclusione è a tratti meno univoca, soprattutto per quanto riguarda le finestre temporali di riferimento. Ciò è dovuto all'esistenza di una condizione piuttosto comune nella fase di sviluppo del linguaggio, riscontrabile tra il 13-17% della popolazione nei primi 3 anni di età, ovvero quella dei bambini "late talkers" (parlatori tardivi) (M. Cristina Caselli e Zoccolotti 2018). Questo termine è stato creato per descrivere la condizione di quei bambini in cui si evidenzia una marcata discrepanza tra la capacità di comprensione e l'effettiva abilità di produzione verbale.

Naturalmente, non tutti coloro che rientrano all'interno di questa finestra temporale svilupperanno in futuro un disturbo del linguaggio. È stato attestato che questa percentuale si riduce al 5-7% entro il quarto anno di età (Tomblin et al. 1997). Per questa ragione questa età assume un'importanza fondamentale ai fini di una corretta diagnosi, svolgendo la funzione di vero e proprio spartiacque. Quei bambini che non recuperano in maniera spontanea nell'età compresa tra i 3-4 anni possono essere diagnosticati con un disturbo primario del linguaggio (Guasti e Marotta 2018).

Nonostante ciò non sono ancora tuttavia chiare quali siano le caratteristiche linguistiche e cognitive da individuare nei bambini late talkers per determinare se tale condizione possa evolversi in un vero e proprio disturbo completo (M. Cristina Caselli e Zoccolotti 2018).

Comprensibilmente, le caratteristiche specifiche di ogni lingua vanno a determinare criticità differenti nella manifestazione del disturbo. Per il caso specifico dell'italiano è possibile affermare che la presenza di difficoltà nell'uso dei pronomi clitici di III persona (lo, la, gli, le) risulta essere solitamente un buon indicatore della presenza di un disturbo, pur dovendo essere innanzitutto appurati i criteri di esclu-

sione espressi nella sezione 1.1.1, dato che questo fenomeno è riscontrabile anche, per esempio, nella sordità (Guasti e Marotta 2018).

Anche problemi nella ripetizione di "non parole" sembra essere caratterizzante, a prescindere dalla quantità di tempo cui si è stati esposti alla lingua o dalla condizione socio-economica (ibid.).

Tutto ciò è valido anche per i bambini cresciuti bilingui, i quali sembrano presentare caratteristiche pressoché identiche ai bambini monolingui. Gli studi condotti finora sembrano infatti confermare che l'esposizione a più lingue durante la crescita non peggiori la condizione del disturbo (Marini 2014) (Marotta e Maria Cristina Caselli 2014) (M. Chiara Levorato e Andrea Marini 2018). Si deve tuttavia far notare che gli studi in merito sono di numero esiguo e sempre effettuati con un campione limitato (ibid.).

1.2 Aree linguistiche

Le aree linguistiche che possono essere compromesse dai disturbi primari del linguaggio sono state già precedentemente elencate nella sezione 1.1.1, ma le problematiche evidenziate non sono necessariamente limitate a una singola area. Si sovrappongono invece spesso tra loro, motivo per cui si è parlato di un disturbo eterogeneo nelle sue manifestazioni piuttosto che di un singolo fenomeno ben delineato (Guasti e Marotta 2018).

Di seguito verrà fornita una breve spiegazione di quali siano le specifiche caratteristiche di ogni singola area.

1.2.1 Fonologia

La fonologia è la branca della linguistica che si occupa dell'organizzazione dei suoni nel sistema linguistico. La sua unità di base è il cosiddetto "fonema", ovvero un qualsiasi suono producibile dall'apparato fonatorio umano che sia dotato di valore distintivo e oppositivo (Berruto e Cerruti 2011).

La sua compromissione rende problematico l'atto comunicativo stesso, impedendo l'associazione tra suoni e significati e mettendo a rischio la comprensione sia

da parte del parlante che da parte del suo interlocutore (Bishop et al. 2017). Come si può facilmente immaginare, quest'area si situa alla base di tutte le altre, ed è quella in cui i sintomi di un DSL sono più evidenti.

1.2.2 Morfosintassi

La morfosintassi descrive la struttura della parola, analizzando gli elementi primari di cui è composta, ossia i "morfemi". Si tratta di unità di prima articolazione che posseggono un significato proprio, con valore e funzione ben determinati e identificabili, nonché riproducibili (Berruto e Cerruti 2011).

Un esempio di problematiche connesse a quest'area è già stato fornito nella sezione 1.1.2, dove si fa menzione delle difficoltà nell'uso dei pronomi clitici in italiano.

1.2.3 Sintassi

La sintassi ha a che fare con le interazioni tra le parole all'interno della frase e la loro capacità di formare strutture che si comportano come una sola di esse. Queste strutture vengono definite "sintagmi" o "costituenti" (ibid.).

Le problematiche che insorgono a questo livello possono ridurre significativamente la produzione di enunciati complessi oppure compromettere l'intelligibilità degli enunciati stessi (Bishop et al. 2017).

1.2.4 Semantica lessicale e frasale

La semantica lessicale e frasale affronta il problema della categorizzazione del significato, o dei significati, che una parola può assumere e delle sue interazioni con le altre parole sia a livello di un singolo enunciato che a livello testuale (Berruto e Cerruti 2011).

In quest'area ciò che viene compromesso sono le competenze del paziente da un punto di vista del vocabolario, sia nella produzione che nella comprensione, e nell'associazione del significante a un significato.

È di frequente riscontro nei bambini affetti da DSL, per esempio, una difficoltà nel ripetere "non parole", già citata nella sezione 1.1.2, abilità fondamentale durante la crescita per l'ampliamento del proprio vocabolario (Gathercole 2006). Quest'ultimo è difatti spesso ristretto rispetto ai pari età, limitando così le possibilità espressive del bambino (Guasti e Marotta 2018).

1.2.5 Pragmatica

La pragmatica descrive il livello più complesso del linguaggio, ovvero il livello di interazione tra il sistema linguistico e i suoi utenti. Si basa sulla nozione di "enunciato", ovvero l'espressione linguistica utilizzata in una situazione specifica il cui valore va ricostruito a partire dalla situazione stessa.

Il valore semantico delle parole che compongono un enunciato non determinano in maniera univoca il valore comunicativo dell'enunciato. Il significato scaturisce pertanto dall'uso che ne fanno i parlanti, adottando convenzioni che hanno effetti concreti sulla realtà contestuale con cui si interfacciano. In tal senso la pragmatica individua quelle "regole" a cui l'espressione degli enunciati è soggetta (Berruto e Cerruti 2011).

Ciò che emerge come problematico a questo livello è spesso l'appropriatezza contestuale delle espressioni utilizzate dai bambini affetti da DSL (Bishop et al. 2017).

1.3 Perché intervenire tempestivamente

Data la possibilità di individuare questo ritardo del linguaggio dei parlatori tardivi già dall'età di 24 mesi, determinare delle caratteristiche linguistiche specifiche che possano essere considerate buoni indicatori di un eventuale insorgenza di un DSL è fondamentale per riuscire a formulare una diagnosi precoce. In tal modo, sarebbe possibile ridurre o anche prevenire le pesanti conseguenze cui i bambini affetti da DSL vanno incontro.

Si osserva che in molti casi questi disturbi portano infatti con sé anche ripercussioni sul piano sociale, emotivo ed educativo, in particolare nel momento di passaggio

dall'oralità alla scrittura, ovvero con l'inizio del percorso scolastico, una situazione che può portare allo sviluppo di ulteriori problematiche nell'area psicologica (Guasti e Marotta 2018).

Numerosi studi hanno inoltre dimostrato che nell'arco della crescita il recupero interessa spesso solo il livello "formale" del linguaggio: le capacità nell'organizzazione testuale e narrativa restano invece di basso livello (Bishop et al. 2017) (Association et al. 2013).

Come spiegato da Caselli e Zoccolotti all'interno del documento prodotto durante la Consensus Conference, le direttive internazionali raccomandano un monitoraggio pediatrico iniziale entro i 36 mesi (M. Cristina Caselli e Zoccolotti 2018):

Per permettere questo è fondamentale però verificare che gli screening condotti siano "efficaci" e cioè: a) siano sensibili e specifici per ciò che "selezionano"; b) utilizzino degli strumenti validati, specifici ecologici ed economici e con un buon valore prospettico; c) si focalizzino su una fascia d'età precoce ma che consideri l'ampia variabilità dell'acquisizione del linguaggio; d) prevedano degli interventi successivi in grado di migliorare l'outcome del linguaggio; e) limitino il più possibile gli effetti "collaterali" (ad es. focus eccessivo su aspetti dello sviluppo che non necessitano di intervento specifico).

A tal fine, in questa tesi si intende considerare una possibile applicazione degli ultimi progressi nel campo dell'informatica, e in particolare del *machine learning*, al campo della logopedia.

Mediante lo sviluppo di un modello predittivo, sarebbe infatti possibile ottenere maggiori informazioni sulla rilevanza degli indici linguistici da tenere in considerazione, nonché fornire uno strumento aggiuntivo per la diagnosi tempestiva di questi disturbi.

Se si vedessero confermati i sospetti evidenziati nella precedente letteratura su questo argomento, si arriverebbe così ad avere delle maggiori certezze su quali siano le specifiche aree linguistiche in cui andare a ricercare le debolezze di un possibile

paziente. Grazie a questa conoscenza, si potrebbe quindi iniziare un percorso di cura precoce, limitando il rischio di un mancato recupero di alcune abilità.

È certamente necessario poter effettuare previsioni con un livello di accuratezza molto alto affinché si possa giungere a tali conclusioni e l'esperimento oggetto di questa tesi intende proporsi solo come un primo passo verso il raggiungimento di un tale traguardo.

Capitolo 2

Strumenti utilizzati

Per l'estrazione e l'analisi dei dati sono stati utilizzati due software.

Innanzitutto è stato utilizzato il software ELAN per la trascrizione ortografica in formato CHAT-LABLITA delle registrazioni ottenute dalle sessioni di terapia logopedica. Dal testo annotato è stato possibile ricavare statistiche per ogni singolo bambino, successivamente utilizzate per l'analisi e la costruzione di un modello predittivo con il *toolkit* Orange.

In questa sezione verranno introdotti entrambi i programmi e verrà spiegato il loro utilizzo.

2.1 ELAN

Le trascrizioni delle sessioni di terapia all'interno dello studio clinico sono state prodotte utilizzando il software *ELAN*, attenendosi al formato CHAT-LABLITA (Cresti e Moneglia 2005), e sono state successivamente analizzate e annotate sotto il profilo lessicale, morfo-sintattico, pragmatico e narrativo (Max Planck Institute for Psycholinguistics 2019) (Brugman, Russel e Nijmegen 2004).

ELAN è un software per l'annotazione di file video e audio. Tale processo consiste nella creazione di una stringa di testo che può contenere una frase, una parola o qualsiasi descrizione di un fenomeno osservato nel file. Le annotazioni vengono create su diversi livelli, che possono essere anche interconnessi da un punto

di vista gerarchico, chiamati "*tier*". Il contenuto testuale è codificato con il sistema standard di codifica Unicode e la trascrizione è salvata in formato XML.

Dai file contenenti le annotazioni è stato possibile ottenere dati statistici sui fenomeni linguistici sovracitati, poi utilizzati come base per l'esperimento di predizione effettuato con Orange.

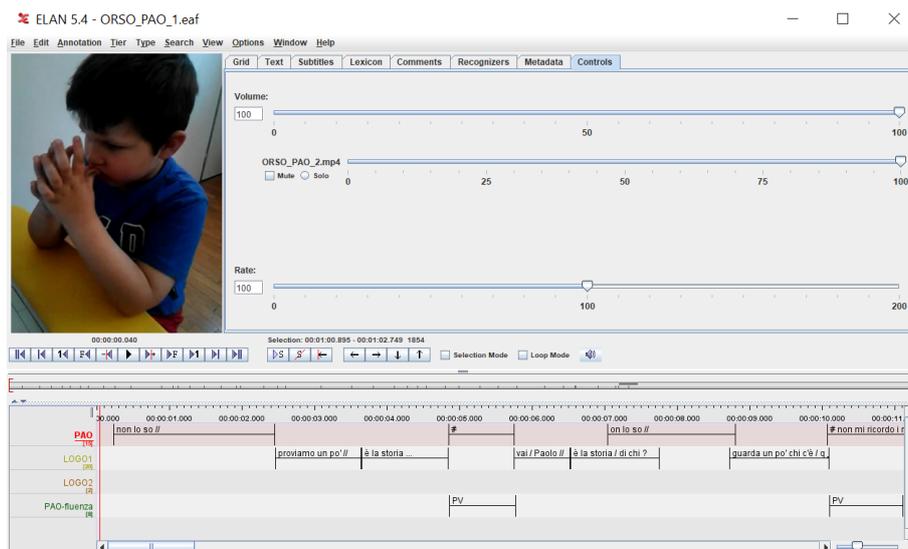


Figura 2.1: Interfaccia di ELAN

2.2 Orange

Orange è un pacchetto software di *machine learning* e *data mining* per l'analisi dei dati basato sul linguaggio di scripting *Python*. Esso sfrutta un linguaggio di programmazione visuale, le cui singole funzioni sono organizzate da un punto di vista gerarchico (Demšar et al. 2013).

I processi di basso livello come il filtraggio dei dati o il calcolo probabilistico possono essere assemblati in algoritmi di livello più alto come un *classification tree*, permettendo così di manipolare facilmente il codice alla base aggiungendo o rimuovendo diversi tipi di funzioni direttamente dall'interfaccia del *workflow*. Per una visualizzazione dei dati interattiva, queste diverse funzioni sono rappresentate sotto forma di icone chiamate "*widget*" (fig. 2.2) (ibid.).

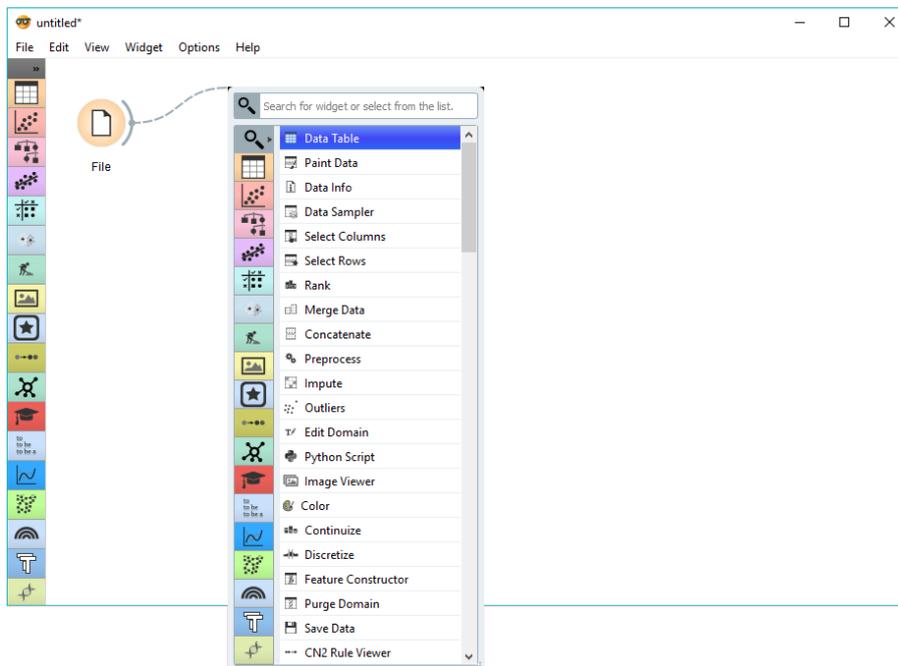


Figura 2.2: Esempio di un *workflow* con il *widget* "File"

Lo sviluppo del *workflow* avviene mediante la connessione dei *widget* tra loro, ognuno soggetto a particolari restrizioni per il tipo di input e output ricevuto.

Tramite il *widget* "File", per esempio, è possibile caricare i propri dati in vari tipi di formati, tra cui anche Excel, e rappresenta il passo fondamentale per la preparazione ad una successiva elaborazione. Il *widget* "File" suddivide infatti in maniera automatica i dati presenti nel documento di origine in "*variabili*" e "*feature*". Ad esso possono essere connessi diversi altri tipi di *widget*, come per esempio "Data Table", utilizzato per la consultazione dei dati, oppure "Select Columns", per la definizione manuale delle variabili e delle feature.

Le funzioni utilizzate ai fini di questa tesi sono suddivise nelle aree seguenti (ibid.):

1. **Data management and preprocessing:** per gli input e gli output dei dati, il loro filtraggio, modifica e selezione.
2. **Classification:** per l'implementazione di svariati tipi di algoritmi di *machine learning*.
3. **Regression:** per l'implementazione di modelli di regressione.

4. **Association:** per l'implementazione di regole associative.
5. **Clustering:** per algoritmi di *clustering* come "k-means".
6. **Evaluation:** con funzioni per la *cross validation* e la valutazione qualitativa dei metodi predittivi.

Capitolo 3

Metodologia

In questo capitolo verrà illustrato innanzitutto lo studio clinico condotto presso l’Azienda USL Toscana Centro da cui sono stati ottenuti i dati linguistici per l’esperimento oggetto di questa tesi. In seguito, verranno descritte le procedure di elaborazione e raffinamento dei dati grezzi per prepararli a essere utilizzati all’interno di Orange.

Infine, verranno descritti i diversi passaggi effettuati durante l’esperimento per sviluppare il *workflow* e testarne le capacità predittive.

3.1 Studio clinico sulle abilità narrative di bambini affetti da DSL

Lo studio parte dalla considerazione che l’ambito narrativo è uno dei punti critici della riabilitazione logopedica, in quanto questo è l’ambito in cui il recupero delle abilità del paziente sembra essere minore. Essendo un’abilità trasversale e fondamentale per ogni tipo di interazione umana, tanto nel personale quanto nell’apprendimento e nel lavoro, un deficit non recuperato in questo ambito può comportare problemi significativi per quanto riguarda la sfera socio-relazionale durante lo sviluppo (Beraldi, Gagliardi e Milvia 2018).

L’obiettivo è quindi quello di valutare le capacità di bambini in età prescolare nell’ambito della narrazione, in particolare tramite compiti di retelling (cioè la

ripetizione in forma orale di un racconto precedentemente proposto).

L'analisi si sofferma sugli aspetti linguistici e pragmatici, oltre che sul medium utilizzato per far apprendere la storia ai bambini (Beraldi, Gagliardi e Milvia 2018).

3.1.1 Metodo

Il campione da cui si è partiti è costituito da 16 bambini monolingui, 13 maschi e 3 femmine, di età compresa tra 4;2 e 5;4. Questi sono divisi in due gruppi, di cui il primo, composto da 8 soggetti maschi, include unicamente pazienti in trattamento presso la AUSL Toscana Centro in seguito a diagnosi di DSL di tipo espressivo; il secondo, composto da 5 soggetti maschi e 3 femmine, rappresenta invece il gruppo di controllo, con uno sviluppo cognitivo e linguistico nella norma (ibid.).

Il test consisteva di tre compiti differenti:

1. Il test "*Bus Story*" (Renfrew 2016), ormai ben consolidato e basato sulla rievocazione verbale e sulle abilità descrittive. Medium: Tavole figurate.
2. retelling de "*I tre porcellini*". Medium: Libro cartaceo illustrato.
3. retelling del cortometraggio di 90 secondi "*Orso Polare*". Medium: Tablet.

Ogni singolo test è stato videoregistrato, in modo da essere poi trascritto ortograficamente con il software ELAN (si veda Capitolo 2, Sezione 2.1) in formato CHAT-LABLITA (Cresti e Moneglia 2005) e analizzato secondo le caratteristiche evidenziate nella Tabella 3.1.

Data la ridotta dimensione dei campioni, al fine di valutare la significatività statistica delle differenze tra i due gruppi, si è deciso di utilizzare il test del χ^2 e il test di Kolmogorov-Smirnov in quanto non parametrici e più adatti considerata la quantità di dati a disposizione.

LIVELLO DI ANALISI	INDICI
Fluenza	fluenza narrativa false partenze pause (vuote/piene)
Tratti lessicali e morfo-sintattici	type/token ratio MLU n. enunciati verbles n. enunciati interrotti n./percentuale di principali, coordinate e subordinate errori di morfologia libera/legata n. (e correttezza) dei clitici coerenza dei tempi verbali
Abilità Narrative	grammatica delle storie
Comunicazione non verbale	Gesto mimica: sorrisi, atteggiamenti di ricerca di conferma e espressioni concordi all'emozione del racconto contatto oculare

Tabella 3.1: Descrizione degli indici linguistici considerati nello studio

3.1.2 Risultati

Lo studio ha evidenziato differenze significative da un punto di vista statistico (con un $p\text{-value} < 0.05$) tra i due gruppi, soprattutto nella fluenza, nei tratti lessicali e morfosintattici e nell'uso di auto- ed etero-adattatori, in accordo con la letteratura sul tema.

I bambini affetti da DSL dimostrano infatti di essere ipofluenti, soprattutto nelle prove non standardizzate, e per entrambi i gruppi le performance ottenute con l'utilizzo del supporto multimediale risultano essere peggiori. Ciò sembra quindi suggerire che le modalità di presentazione della storia possano avere un effetto concreto sull'abilità narrativa.

Sotto il punto di vista lessicale e morfosintattico, gli enunciati prodotti dai bambini affetti da DSL sono più frequentemente interrotti, mentre si riscontra anche una forte discrepanza nell'uso di strutture frastiche complesse e dei pronomi clitici, questi ultimi il più delle volte assenti.

A ciò si aggiunge un utilizzo maggiore di gesti non connessi al discorso da parte dei bambini con DSL, interpretata dalle autrici come possibile espressione di disagio nell'esecuzione del task (Beraldi, Gagliardi e Milvia 2018).

3.2 Preparazione dei dati

I dati grezzi ottenuti durante lo studio condotto presso la AUSL Toscana Centro sono stati raccolti all'interno di file Excel separati. Tuttavia, per l'esperimento oggetto di questa tesi è stato necessario innanzitutto unire tutti i file in una singola tabella, in modo che il programma potesse avere a disposizione tutti i dati nella maniera più immediata possibile, senza dover ricorrere a file multipli.

Per fare ciò sono state apportate delle modifiche al contenuto dei file, in merito alle descrizioni dei dati. Sono state aggiunte le informazioni che esplicitano ogni colonna a quale delle tre prove faccia riferimento. Inoltre, tutti gli spazi prima presenti nei metadati sono stati sostituiti con il carattere "_". È stato infatti riscontrato che talvolta il carattere di spazio può causare problemi durante la lettura da parte del programma Orange.

L'ultimo passaggio per preparare i dati è stato effettuato direttamente nell'interfaccia di Orange. Caricati i dati tramite il *widget* "File", infatti, il programma riconosce automaticamente le colonne in cui sono inseriti i metadati e quelle dove sono presenti le feature con la propria tipologia. Nel caso di questo esperimento alcune delle feature sono state erroneamente definite come variabili, quindi è stato utilizzato il *widget* "Select Columns" per individuare manualmente le variabili desiderate per l'esperimento. Si tratta della categoria "CLASSE" contenente informazioni sull'oggetto: "DSL" oppure "Normotipo".

3.3 Esperimento con Orange

Eseguite queste correzioni nel contenuto dei dati, è stato possibile procedere con l'esperimento. Di seguito viene fatta un'introduzione ai *widget* utilizzati, vengono spiegate le scelte compiute e come questi widget sono stati utilizzati nel contesto dell'esperimento.

3.3.1 L'algoritmo "kNN"

L'algoritmo prescelto da allenare è "*k-nearest neighbors*" (kNN), molto utilizzato nel *machine learning* per risolvere problemi di classificazione e regressione (Demšar et al. 2013). Si tratta di un algoritmo di apprendimento supervisionato che parte da un insieme di esempi di oggetti noti utilizzati come base per classificare nuovi tipi di oggetti. Per fare ciò, deve quindi avere a disposizione:

1. Un *training set*;
2. Una metrica per determinare la distanza tra gli oggetti;
3. Il valore di k .

k è definito come il numero di "*neighbor*" o "vicini" del nuovo oggetto da utilizzare per la classificazione. I "*neighbor*" sono gli oggetti che, nel *training set*, presentano la minore distanza dall'oggetto sconosciuto, ovvero i cui dati sono maggiormente simili ad esso.

La scelta del corretto valore di k è essenziale poiché, se questo fosse troppo piccolo, allora l'algoritmo risulterebbe essere molto sensibile al rumore, mentre se fosse troppo grande si rischierebbe di includere oggetti appartenenti ad altre classi. L'algoritmo kNN calcola la media dei valori di questi "*neighbor*" e la utilizza per predire la classe del nuovo oggetto. In Orange il *widget* corrispondente è denominato "kNN" e come output fornisce un "*Learner*", utilizzato nel test di classificazione, e un "*Model*", utilizzato nel test di regressione.

Per l'esperimento in oggetto il training set è costituito dai dati raccolti e raffinati come spiegato nella sezione 3.2. La metrica prescelta è "Euclidea", ovvero la semplice distanza tra due punti.

Data la dimensione ridotta del campione, si è optato per il valore $k = 3$.

3.3.2 Classificazione

Per il compito di classificazione è necessario aggiungere un altro *widget* al *workflow* di Orange. Questo *widget* si chiama "*Test & Score*" (fig. 3.1) e serve a valutare la capacità di predizione degli algoritmi di *machine learning*, come per esempio il

"kNN" (Demšar et al. 2013). Come input deve ricevere i dati dal "*Learner*" generato da un *widget* presente sul *workflow* e deve avere un *dataset* di riferimento. Utilizza diversi tipi di metodi per valutare la performance dei classificatori e come output genera i valori "*Recall*", "*Precision*" e "*F-Score*" (abbreviato in "F1").

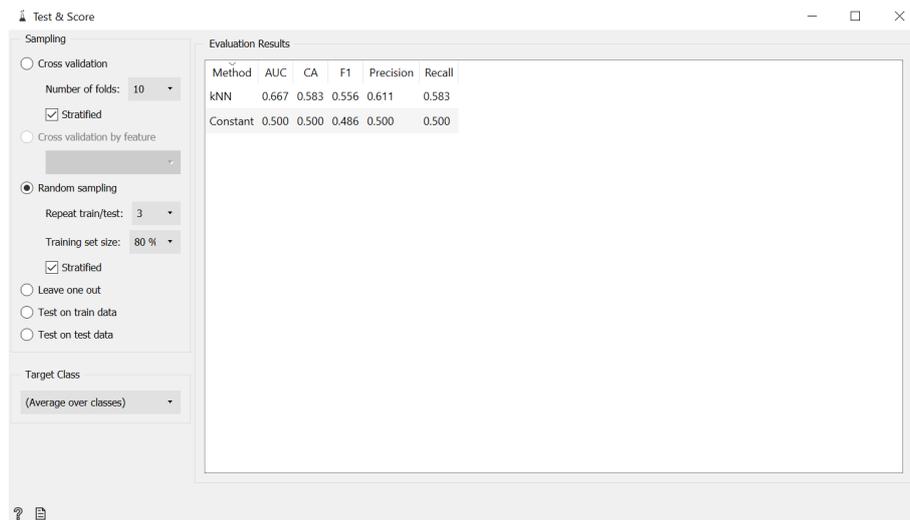


Figura 3.1: Interfaccia *widget* "*Test & Score*"

Il valore di "*Recall*" esplicita il rapporto tra il numero di oggetti recuperati dal sistema e il totale da recuperare se lo stesso funzionasse alla perfezione. Questo può assumere tutti i valori $0 \leq x \leq 1$, con 1 che sta a determinare il valore massimo per cui il sistema ha recuperato tutti gli oggetti (Powers 2011).

Il valore di "*Precision*" esplicita il rapporto tra il numero di oggetti correttamente recuperati dal sistema e il totale degli oggetti recuperati inclusi gli errori. Determina cioè il tasso di errore ed esprime il tasso di probabilità che un elemento recuperato sia stato classificato correttamente. Anch'esso può assumere tutti i valori $0 \leq x \leq 1$ (ibid.).

Il valore di "*F-Score*" rappresenta la media armonica di *Precision* e *Recall*. I valori che può assumere sono chiaramente anch'essi $0 \leq x \leq 1$ e maggiore è la vicinanza al valore massimo "1", maggiore sarà l'accuratezza del test (ibid.).

Gli output di questo *widget* sono quindi determinanti ai fini della riuscita dell'esperimento.

Nel caso dell'esperimento in questione sono stati collegati al "Test & Score" i widget "Select Columns" con il dataset generato nella sezione 3.2 e il widget "kNN" (fig. 3.2). È stato utilizzato il metodo di campionamento per randomizzazione, in cui il *Learner* viene allenato solo su una percentuale degli oggetti nel *dataset*, per poi essere testato sui rimanenti. La dimensione del *training set* è stata impostata all'80% e il test è stato ripetuto 3 volte.

Il programma ha generato come output un valore di *Precision* pari a 0,611, un valore di *Recall* pari a 0,583 e un valore di *F-Score* pari a 0,556.

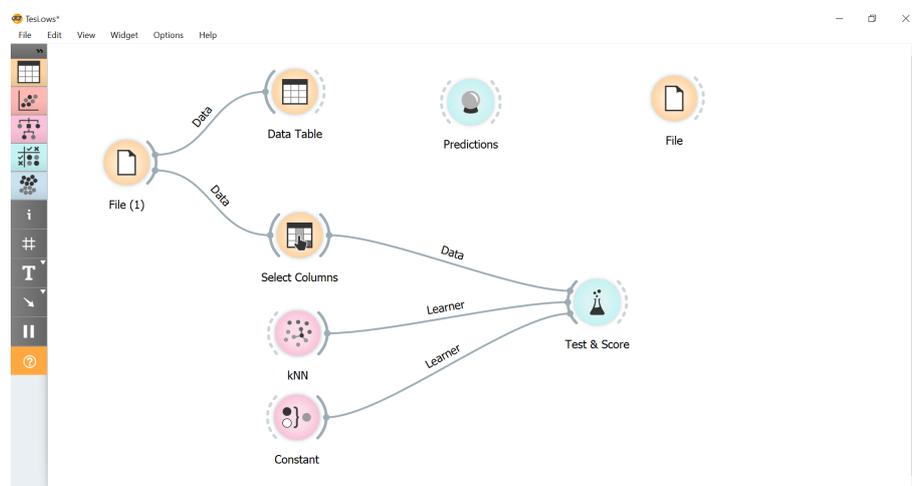


Figura 3.2: Workflow per il test di classificazione

Nella figura 3.1 si evidenzia anche il confronto con i risultati ottenuti dal widget "Constant", il quale predice sempre la classe di maggioranza presente nel *dataset* e viene per questo utilizzato come *baseline* per gli altri modelli. Come si può vedere, infatti, il suo output è un valore di *Precision* pari a 0,500, un valore di *Recall* pari a 0,500 e un valore di *F-Score* pari a 0,486 (ibid.).

3.3.3 Regressione

Il test di regressione viene utilizzato per valutare l'effettiva capacità predittiva del modello sviluppato e per attestare il corretto svolgimento del workflow. Per fare ciò si utilizza il widget "Predictions", come si può osservare nella figura 3.3.

Questo widget riceve come input un *dataset* e un "*predictive model*" (un modello predittivo). Come output produce una lista di dati con annesse le predizioni

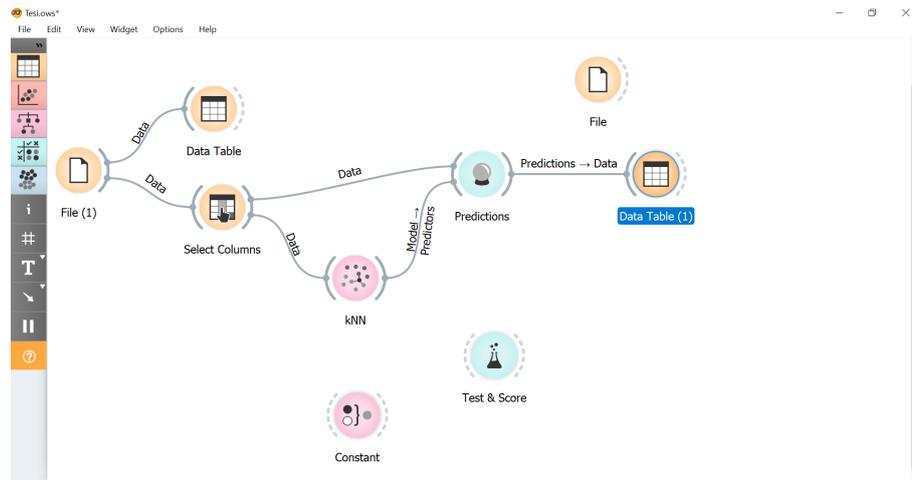


Figura 3.3: Workflow per il test di regressione

effettuate. L'interfaccia del *widget* mostra le probabilità e le predizioni compiute dai modelli. I risultati possono essere osservati utilizzando il *widget* "Data Table".

Come si può vedere nella figura 3.4, il modello ha commesso un singolo errore di predizione nel caso del gruppo di bambini normotipo, mentre ne ha commessi tre nel caso dei bambini affetti da DSL.

	CLASSE	SOGG	kNN	kNN (DSL)	kNN (NORMO)	n_i
1	DSL	ANT	DSL	1.000	0.000	2.0
2	DSL	DAV	NORMO	0.333	0.667	6.0
3	DSL	EDD	DSL	0.667	0.333	0.0
4	DSL	FRA	NORMO	0.333	0.667	4.0
5	DSL	LEO	DSL	0.667	0.333	1.0
6	DSL	LOR	NORMO	0.333	0.667	2.0
7	DSL	NIC	DSL	0.667	0.333	0.0
8	DSL	PAO	DSL	0.667	0.333	3.0
9	NORMO	CHI	NORMO	0.333	0.667	3.0
10	NORMO	COS	NORMO	0.000	1.000	6.0
11	NORMO	DUC	NORMO	0.333	0.667	7.0
12	NORMO	FLO	NORMO	0.333	0.667	12.0
13	NORMO	MATTE	NORMO	0.333	0.667	3.0
14	NORMO	MATTI	NORMO	0.333	0.667	1.0
15	NORMO	MIR	NORMO	0.333	0.667	9.0
16	NORMO	SOF	DSL	0.667	0.333	0.0

Figura 3.4: Predizioni effettuate con kNN

3.3.4 Risultati

Le predizioni effettuate dal modello sviluppato con Orange sembrano essere ancora imprecise, dati i valori di *Precision* e *Recall* riportati nella sezione 3.3.2. Questi sembrano non essere sufficienti per poter affidare il compito di discriminare quale soggetto possa essere diagnosticato come affetto da DSL. Ciò è confermato anche dai risultati riportati nella sezione 3.3.3.

Una tra le ragioni per cui il sistema non sembra essere in grado di discriminare con sufficiente accuratezza è la possibile inefficacia di alcuni dei parametri linguistici analizzati. Difatti, non tutti i dati forniti hanno ottenuto un *p-value* sufficiente a determinare la loro effettiva significatività statistica (sezione 3.1.2).

Tuttavia, tramite il test di classificazione si sono comunque ottenuti risultati migliori rispetto al semplice modello prodotto da "*Constant*", suggerendo quindi che sia possibile sviluppare un modello predittivo sulla base di questi dati.

L'altra possibile ragione è la quantità limitata dei dati a disposizione del sistema per la fase di allenamento. Pur essendo gli indici linguistici abbastanza numerosi, con soli 16 soggetti il campione da cui sono stati estratti questi dati resta pur sempre di piccole dimensioni.

Date queste considerazioni, i risultati ottenuti sembrano offrire in ogni caso buone speranze di riuscire a migliorare la performance del sistema in futuro. Per gli sviluppi futuri, ulteriori studi potrebbero pertanto concentrarsi sull'aumentare la quantità di dati a disposizione al fine di ripetere l'esperimento e migliorarne i risultati. In particolare si evidenzia la possibilità di ampliare il campione di base, oltre a considerare possibili indici linguistici aggiuntivi.

Conclusioni

Questa tesi si è proposta di offrire una nuova prospettiva su un campo complesso come quello dei Disturbi del Linguaggio. Inizialmente è stato analizzato il problema da un punto di vista teorico, facendo riferimento alla ricca letteratura sul tema.

Nel primo capitolo è stata offerta una definizione di "Disturbo del Linguaggio", di cui sono state esaminate le sfaccettature e le problematiche diagnostiche, ancora ampiamente discusse nel mondo scientifico. Si è prestata particolare attenzione ai criteri di esclusione e inclusione, nonché alle aree linguistiche che risultano più comunemente compromesse.

Sono stati inoltre analizzati i numerosi vantaggi connessi a una diagnosi quanto più tempestiva possibile del disturbo, che ha spinto ad adoperare le più recenti tecniche di *machine learning* e *data mining* nell'ambito dell'esperimento oggetto di questa tesi.

Nel secondo capitolo è stata svolta un'introduzione agli strumenti informatici utilizzati sia nello studio clinico da cui sono stati estratti i dati per l'esperimento che nel contesto dell'esperimento stesso, ovvero il software Elan e il *toolkit* Orange.

Nel terzo capitolo è stato descritto nel dettaglio lo studio svolto presso la AUSL Toscana Centro, con un'attenzione particolare alle metodologie d'indagine e i risultati ottenuti. In conclusione è stata presentata un'applicazione pratica delle ultime tecniche di *machine learning* e *data mining* al campo della logopedia, esaminando l'esperimento oggetto di questa tesi. È stata svolta una descrizione dei vari passaggi sia sotto il profilo teorico che pratico e sono state esplicitate le scelte compiute.

Il modello predittivo sviluppato, nonostante non sia risultato essere totalmente affidabile, sembra offrire buone prospettive per gli sviluppi futuri, aprendo così la strada a ulteriori studi in questo ambito.

Bibliografia

- Association, American Psychiatric et al. (2013). “Diagnostic and statistical manual of mental disorders”. In: *BMC Med* 17.
- Beraldi, Francesca, Gloria Gagliardi e Innocenti Milvia (2018). “Abilità narrative nei bambini con DSL in età prescolare: un’analisi dei comportamenti linguistici e non-verbali nei compiti di retelling”. In: *Proceedings of the 15th Conference of the Italian Association for Cognitive Sciences (AISC 2018)*. A cura di Cristiano Chesi e Giulia Piredda, pp. 3–4.
- Berruto, Gaetano e Massimo Simone Cerruti (2011). *La linguistica. Un corso introduttivo*. UTET De Agostini.
- Bishop, Dorothy VM et al. (2017). “Phase 2 of CATALISE: A multinational and multidisciplinary Delphi consensus study of problems with language development: Terminology”. In: *Journal of Child Psychology and Psychiatry* 58.10, pp. 1068–1080.
- Brugman, Hennie, Albert Russel e Xd Nijmegen (2004). “Annotating Multi-media/Multi-modal Resources with ELAN.” In: *LREC*.
- Caselli, M. Cristina e Pierluigi Zoccolotti (2018). “Inquadramento epidemiologico, fattori di rischio, prognosi, strumenti di valutazione”. In: *Consensus Conference*. (28–29 nov. 2018). Roma: CLASTA e FLI, pp. 6–11.
- Cresti, Emanuela e Massimo Moneglia (2005). *C-ORAL-ROM: integrated reference corpora for spoken romance languages*. Vol. 15. John Benjamins Publishing.
- Demšar, Janez et al. (2013). “Orange: Data Mining Toolbox in Python”. In: *Journal of Machine Learning Research* 14, pp. 2349–2353. URL: <http://jmlr.org/papers/v14/demsar13a.html>.

- Gathercole, Susan E (2006). “Nonword repetition and word learning: The nature of the relationship”. In: *Applied psycholinguistics* 27.4, pp. 513–543.
- Guasti, M. Teresa e Luigi Marotta (2018). “Definizione e Diagnosi del Disturbo del Linguaggio”. In: *Consensus Conference*. (28–29 nov. 2018). Roma: CLASTA e FLI, pp. 2–6.
- Lavorato, M. Chiara e Andrea Marini (2018). “Il Disturbo Primario del Linguaggio nel bilinguismo”. In: *Consensus Conference*. (28–29 nov. 2018). Roma: CLASTA e FLI, pp. 16–20.
- Lavorato, M Chiara e Tiziana Rossetto, cur. (2018). *Consensus Conference sui Disturbi del Linguaggio in età evolutiva*. (28–29 nov. 2018). Roma: CLASTA e FLI.
- Marini, A (2014). “Caratteristiche della elaborazione linguistica in bambini bilingui con disturbi dello sviluppo linguistico”. In: *I Disturbi del Linguaggio*, pp. 66–82.
- Marotta, Luigi e Maria Cristina Caselli (2014). *I disturbi del linguaggio: Caratteristiche, valutazione, trattamento*. Edizioni Centro Studi Erickson.
- Max Planck Institute for Psycholinguistics The Language Archive Nijmegen, The Netherlands (2019). *ELAN (Version 5.7) [Computer software]*. url: <https://tla.mpi.nl/tools/tla-tools/elan/>.
- Powers, David Martin (2011). “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation”. In:
- Renfrew, Catherine E (2016). *Bus Story Test: A test of narrative speech*. Speechmark.
- Tomblin, J Bruce et al. (1997). “Prevalence of specific language impairment in kindergarten children”. In: *Journal of speech, language, and hearing research* 40.6, pp. 1245–1260.