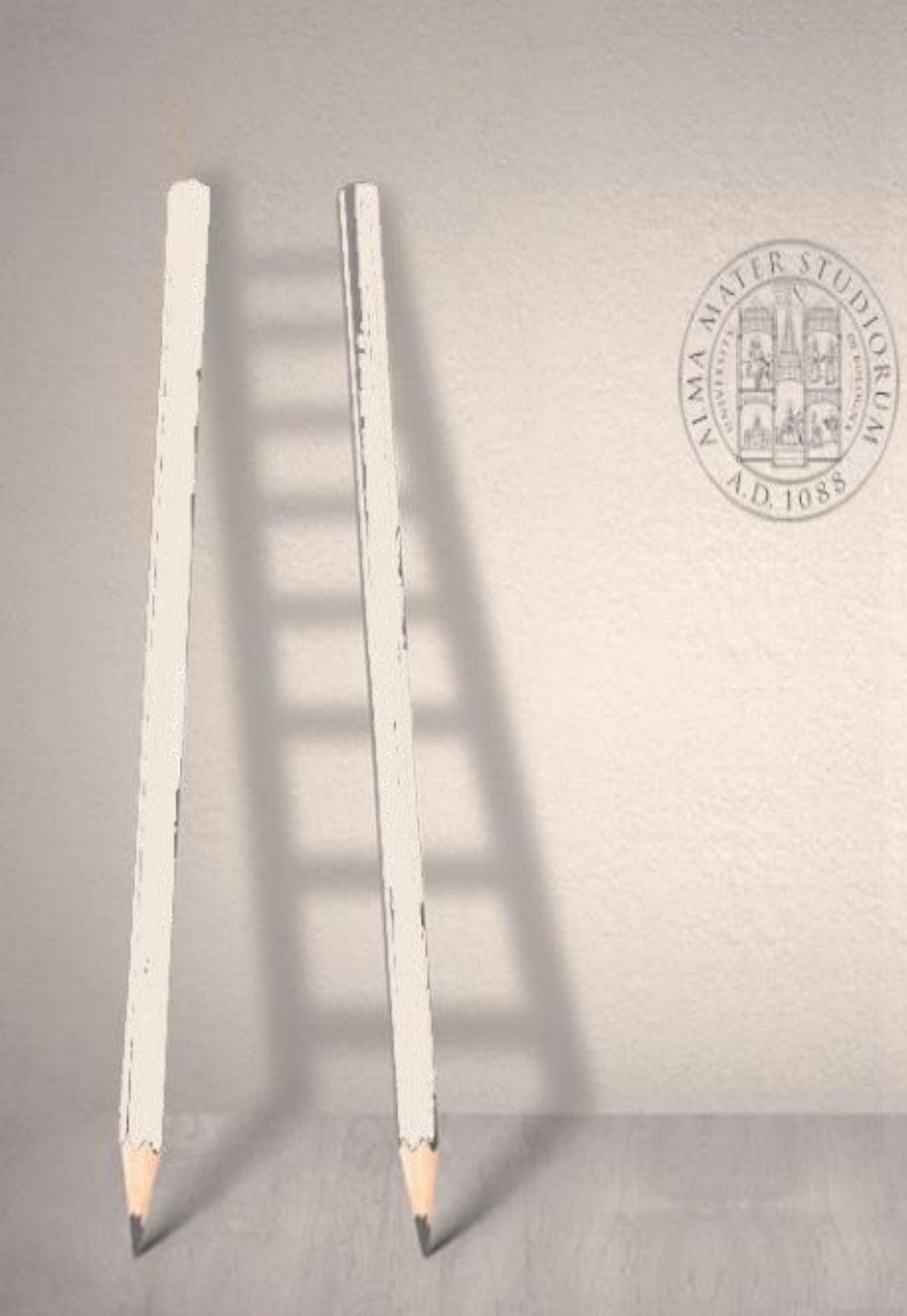


Corpora e contatto. La costruzione di un corpus italiano-ladino

Ilaria Fiorentini – LILEC (Università di Bologna)

LEADhoC Project

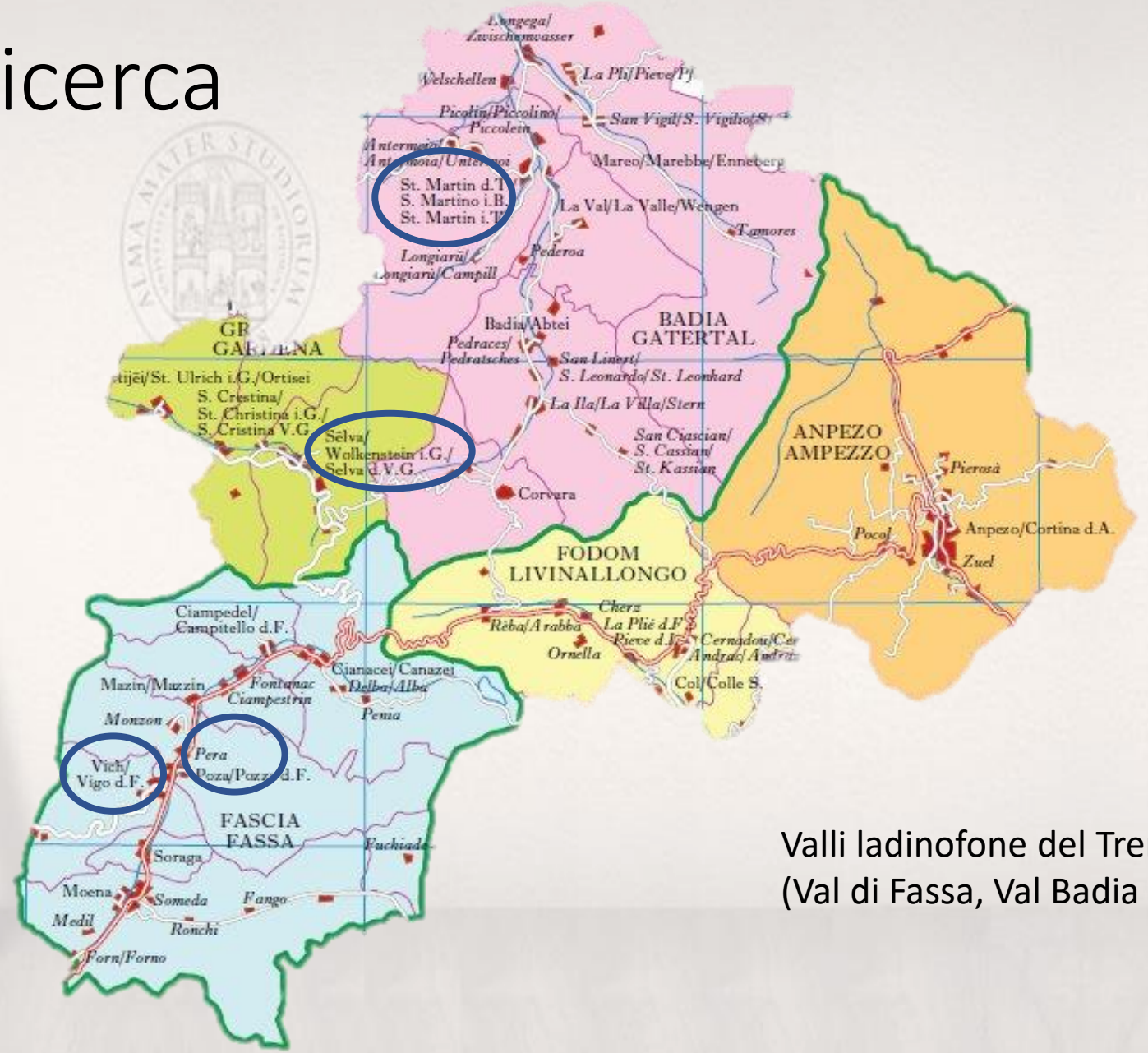
Obiettivi della ricerca

- Indagare la presenza di segnali discorsivi (SD) italiani nel ladino parlato;
- Ricostruire il sistema di SD del ladino.

→Necessità di un corpus di parlato bilingue (italiano-ladino);

→Raccolta di dati di parlato semi-spontaneo (interviste non strutturate) e spontaneo (conversazioni).

Area di ricerca



Valli ladinofone del Trentino Alto-Adige
(Val di Fassa, Val Badia e Val Gardena)

Costruzione del corpus

- Raccolta dati svolta tra aprile 2012 e giugno 2013;
- Selezione degli informanti a partire da contatti interni alle sedi degli Istituti Ladini;
- Rete successivamente allargata a conoscenti o volontari, per avere un campione il più possibile bilanciato (privilegiando la Val di Fassa) per età e genere, variabili ritenute di maggiore interesse per la ricerca;
- Altre variabili considerate: luogo di nascita, luogo di residenza, luogo di nascita di entrambi i genitori, titolo di studio, attività lavorativa, lingua madre.

Struttura del corpus

Valle	Parlanti	Ore	Parole
Fassa	37	8,5	38.634
Gardena	12	2,5	13.282
Badia	8	1,5	10.330
Totale	**57**	**12,5**	**62.246**

Trascrizione e annotazione

- Trascrizione ortografica (sistema ortografico standard di ogni varietà) tramite ELAN;
- Convenzioni impiegate:

  : allungamento vocalico

  | autocorrezione

  / pausa < o = 1 secondo

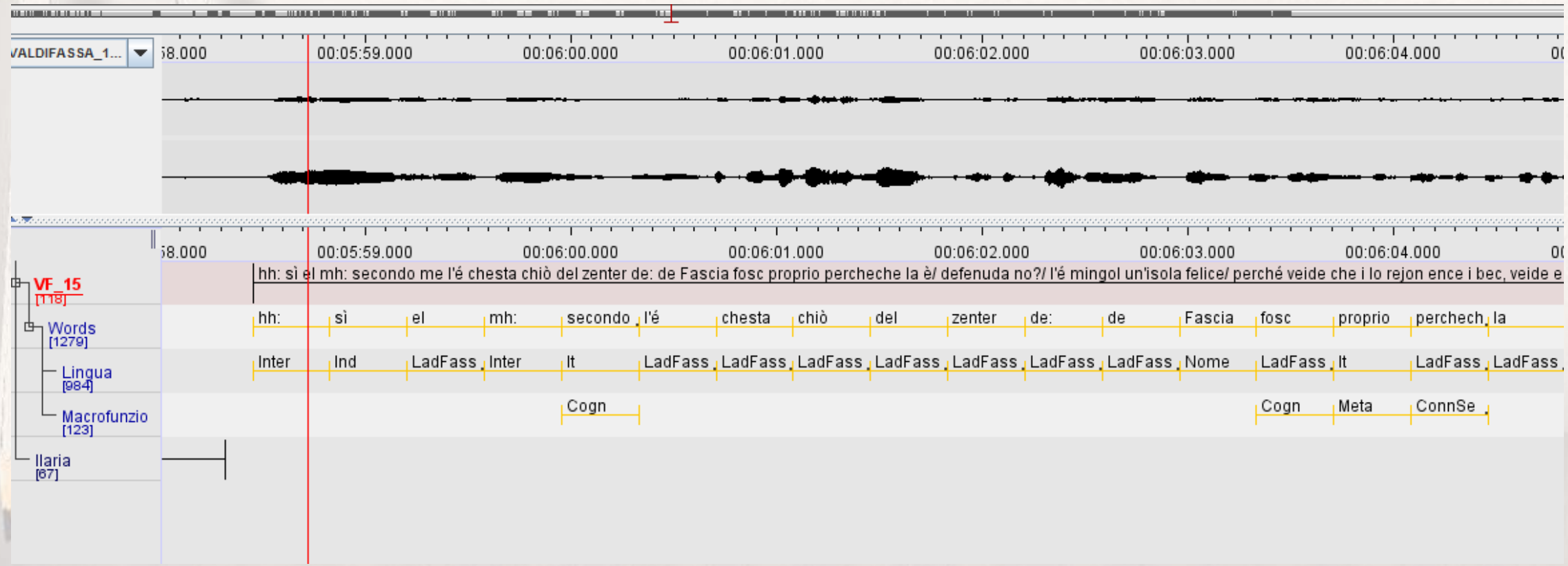
  // pausa > 1 secondo

  = turni sovrapposti

- Successiva tokenizzazione automatica (con revisione manuale) e annotazione dei tokens per lingua e per macrofunzione dei SD.

Un esempio di trascrizione e annotazione

(1) secondo me *l'é: chesta chiò del zenter de: de Fascia fosc* proprio *percheche la è / defenuda no? / l'è: mingol* un'isola felice / perché *veide che i lo rejona ence i bec*



Prospettive future

- Ampliare il corpus in modo da avere più dati di tutte le varietà di ladino indagate;
- Integrare il corpus all'interno del progetto LEADhoC;
- Rendere il corpus pubblico e accessibile;
- …Altre proposte?

Grazie!

Thank you!

Develpai!

Dilan!

De gra!