UNIVERSITÀ DI PISA

Corso di Laurea in Informatica Umanistica

RELAZIONE

Sviluppo di un Sense Inventory per task di Word Sense Disambiguation: il progetto ELEXIS

Candidato: *Francesca Poli*

Relatore: *Alessandro Lenci*

Correlatore: *Monica Monachini, Francesca Frontini*

Anno Accademico 2019-2020

Indice

Introduzione **3**

1 Contestualizzazione del lavoro **6**

1.1 Il progetto ELEXIS e la e-lexicography 6

1.2 Risorse lessicali di partenza 8

2.1.1 Parole Simple Clips 8

2.1.2 ItalWordNet 9

2.1.3 Database di mapping IWN-PSC 10

1.3 Perché un Sense Inventory? 11

1.4 Sviluppo del progetto 12

1.5 Introduzione al Word Sense Disambiguation 13

2 La creazione del Sense Inventory **15**

2.1 Il dataset annotato CoNLL-U 15

2.2 *Parsing* dei dati iniziali 17

2.3 Il formato finale del Sense Inventory 20

2.4 Considerazioni sulla performance del Sense Inventory 23

3 Estensione automatica del mapping di sensi **26**

3.1 Soglia con lemmi mappati 27

3.2 Allineamento di lemmi privi di mapping 29

3.3 Risultati del test di mapping 31

4 Valutazione delle risorse **34**

4.1 L'annotazione semantica manuale 34

4.1.1 Algoritmo di selezione delle frasi per il *test set* 36

4.2 Analisi comparativa della copertura delle risorse 37

5 Conclusioni..... 41

Bibliografia 44

Risorse..... 45

Appendice 46

Introduzione

In questa tesi verrà presentato lo sviluppo di un Sense Inventory per un task di Word Sense Disambiguation inserito nel progetto europeo ELEXIS[1], il cui obiettivo è la costruzione di un'infrastruttura per la lessicografia computazionale. Lo scopo del presente lavoro è di ottenere un inventario di sensi legati ai lemmi estratti da un dataset, una risorsa creata a partire da un corpus parallelo di ELEXIS. Il Sense Inventory verrà poi usato per compiti di annotazione semantica di corpora multilingue, sempre nell'ambito di ELEXIS.

I lessici computazionali e i dizionari elettronici sono risorse linguistiche indispensabili, al contempo oggetto di studio e strumento fondamentale nell'ambito lessicografico digitale. I dati lessicali in ogni formato entrano nel lavoro di ricerca (ma non solo) in una vasta gamma di discipline: per esempio, nel campo prettamente umanistico, i dizionari sono studiati in termini di valori culturali e ideologici che essi contengono, o per il loro ruolo nella standardizzazione della lingua, per dare un'idea della vastità d'informazione che possono veicolare. Nel campo delle tecnologie del linguaggio, acquisiscono un'ancora maggiore rilevanza, dal momento che la ricerca e l'innovazione nel settore si sta sviluppando rapidamente e sta prendendo piede in numerose applicazioni. La linguistica e la lessicografia digitale si avvalgono sempre più spesso di strumenti di analisi quantitativa e di algoritmi di parsing per esplorare la struttura del linguaggio nelle sue regolarità, al fine di trovare nuovi sviluppi per le tecniche di annotazione dei testi. Diventa così necessario costruire risorse base per questo lavoro, come un Sense Inventory, che aiutino nello sviluppo di corpora testuali sempre più riccamente annotati con informazioni su più livelli, per velocizzare e automatizzare per quanto possibile i meccanismi di annotazione, ridurre i tempi e i costi di sviluppo, sviluppare rappresentazioni standardizzate (Lenci *et al.*, 2005). Le risorse lessicali, in particolare i Sense Inventories, diventano elementi fondamentali per lo sviluppo dei metodi di annotazione automatica, per i task di *processing* del linguaggio naturale, per i calcoli di linguistica computazionale e per l'analisi semantica di *big data*. In questo ampio contesto si colloca il lavoro svolto e presentato in questa tesi.

I lessici obbligano ad una continua manutenzione e integrazione per essere aggiornati e ampliati seguendo le esigenze della ricerca e dell'utenza:

[1] Sito web ufficiale del progetto. https://elex.is/

> « Possedere informazioni affidabili e accurate sul significato e l'uso delle parole è cruciale nella società multilingue e multiculturale di oggi. Tradizionalmente, queste informazioni si trovavano nei dizionari - monolingue, bilingue o multilingue. I dizionari cartacei stanno andando lentamente in disuso, ma i database dei dizionari sono ora integrati in siti web, applicazioni mobili, prodotti e servizi digitali.» [2]

I corpora di diversa natura, strutturazione e compilazione diventano le fondamenta del passaggio al nuovo modello di e-lexicography. La loro affidabilità e copertura dev'essere costantemente monitorata e migliorata, così come la loro accessibilità e compatibilità nel formato, che dev'essere sempre armonizzato al contesto in cui verrà inserito il corpus.

In questo elaborato descriveremo il Sense Inventory in italiano creato, come progetto di tesi, nel contesto di ELEXIS, analizzando le basi lessicografiche su cui è stato costruito, lo scopo, gli strumenti e i metodi tramite i quali è stato sviluppato. Inoltre, saranno presentate le valutazioni dei risultati e saranno descritti gli altri task correlati al Sense Inventory: nello specifico, il calcolo automatico della similarità tra sensi delle risorse e la creazione di un test-set per valutare sia la copertura dei lessici sia la performance degli algoritmi sviluppati per l'automatizzazione dei processi necessari per il progetto, parte prettamente computazionale di questa tesi.

La tesi si suddivide in quattro capitoli: un'introduzione al progetto ELEXIS e al contesto di lavoro; la descrizione e analisi dei processi per la creazione del Sense Inventory, conforme agli standard richiesti da ELEXIS; l'esposizione delle modalità di estensione del mapping dei sensi del Sense Inventory provenienti da fonti lessicali differenti; infine, un resoconto delle valutazioni e validazioni dei dati utilizzati e della risorsa lessicale, oltre ad una sezione dedicata al task di annotazione semantica manuale. Saranno quindi approfondite le modalità di svolgimento delle diverse fasi di lavoro, i cui risultati saranno riportati nelle conclusioni finali.

La scelta di questa tematica per la tesi nasce dalla portata innovativa, dal respiro internazionale e dall'intrinseca interdisciplinarità del progetto, che richiama collaboratori, ricercatori e anche utenti e fruitori provenienti da numerosi campi di studio: umanisti, ingegneri del linguaggio, linguisti e lessicografi, scienziati, con competenze tecniche più o meno avanzate. Questa contaminazione e interscambio di domini e conoscenze è proprio quella che caratterizza il Corso di Studi in Informatica Umanistica e che rende il progetto ELEXIS così interessante e stimolante per una tesi di laurea.

[2] Sezione "obiettivi" del sito di ELEXIS. https://elex.is/objectives/

1 Contestualizzazione del lavoro

I dizionari sono gli strumenti scientifici per eccellenza: per questo c'è bisogno di un'infrastruttura lessicografica forte e robusta, che a sua volta necessita di grande impegno e lavoro.

Gli strumenti sviluppati all'interno del progetto ELEXIS si propongono come pionieri di un nuovo modo di sviluppare e trattare dati lessicografici che siano di alta qualità, open-source e utilizzabili da ricercatori, istituzioni e molti altri settori. ELEXIS punta alla creazione di una piattaforma comune di conoscenze e competenze lessicografiche e di linguistica computazionale, con risorse multilingui e multifunzionali, che porti ad una vera interdisciplinarietà produttiva, avanzata ed efficiente [10].

Ecco perché la realizzazione di risorse e di strumenti di analisi di dati, come un Sense Inventory o un algoritmo per il calcolo della similarità tra sensi per il mapping, acquisiscono particolare rilevanza all'interno di un progetto così articolato; allo stesso modo, è per questo fondamentale il monitoraggio dello stato dell'arte delle risorse già esistenti e in uso. Questo capitolo vuole fornire il contesto di lavoro della tesi, a partire da tali considerazioni iniziali.

1.1 Il progetto ELEXIS e la e-lexicography

Il progetto ELEXIS ha origine dall'ENeL (European Network of e-Lexicography), che tra il 2013 e il 2017 ha dato vita all'iniziativa COST (European Cooperation in Science and Technology) [3], un framework intergovernativo paneuropeo, la cui missione è di permettere sviluppi scientifici e tecnologici rivoluzionari che portino a nuovi concetti e prodotti per nutrire la ricerca e l'innovazione in Europa [10].

La lessicografia come campo ha una lunga tradizione di perfezionamento della descrizione semantica delle singole lingue in dizionari monolingui completi o di analisi contrastiva dettagliata tra due o più lingue in dizionari bilingue e multilingue. Tuttavia, queste risorse non sono attualmente utilizzate nell'ambito delle tecnologie linguistiche esistenti ed emergenti. Sono quasi completamente assenti nei cloud di dati open-source e nelle tecnologie del Semantic Web, e sono di fatto "digitalmente invisibili". Nell'ultimo decennio è emerso il nuovo campo della e-

[3] Iniziativa dell'ENeL COST. www.elexicography.eu

lexicography, che può essere visto in iniziative come appunto l'ENeL COST, la serie di conferenze eLex [4] o il workshop Globalex a LREC 2016[5] e 2018[6].

L'innovativo campo dell'e-lexicography è dedicato alla creazione di dizionari digitali definiti come risorse lessicografiche destinate agli utenti umani, ma che si distaccano dal supporto cartaceo per esplorare le possibilità quasi infinite del nuovo ambiente digitale, con l'obiettivo di portare la ricerca e fruizione lessicografica a livelli completamente diversi. In questo contesto, l'apprendimento automatico, il data mining e altre tecniche computazionali stanno iniziando a trovare la loro strada nella lessicografia. Combinando sia le conoscenze e le competenze lessicografiche tradizionali che la linguistica computazionale, e coinvolgendo nel processo anche comunità linguistiche più ampie, si crea un enorme potenziale per lo sviluppo del settore.

ELEXIS propone un'organizzazione inclusiva a più livelli, che coinvolga diversi gruppi di ricerca con diverso background e favorisca la cooperazione e lo scambio di conoscenze tra le diverse comunità di ricerca in lessicografia, al fine di colmare il divario tra le lingue con minori risorse e quelle con un'esperienza avanzata di e-lexicography. Lo scopo dell'organizzazione è di dare vita ad un nuovo tipo di lessicografia che non consideri più le lingue come entità isolate, ma abbracci pienamente la natura paneuropea delle lingue parlate in Europa e si estenda in futuro anche a livello globale.

L'infrastruttura di Ricerca Europea "Common Language Resources and Technology Infrastructure" (CLARIN)[7] è un partner fondamentale di ELEXIS. CLARIN ha come scopo principale quello di rendere tutte le risorse e gli strumenti linguistici digitali da tutta Europa e oltre accessibili attraverso un unico ambiente online, per il supporto dei ricercatori nelle scienze umane e sociali; pertanto, è pienamente iscritto e vicino agli ideali e scopi di ELEXIS. CLARIN supporta il progetto ELEXIS, fornendo dati e piattaforme di lavoro in collaborazione con un'altra infrastruttura europea, la Digital Research Infrastructure for the Arts and Humanities (DARIAH), "un ecosistema d'avanguardia per la ricerca umanistica europea" [8].

[4] Conferenze eLex. https://elex.link/
[5] Workshop Globalex a LREC 2016. https://globalex2016.globalex.link/
[6] Workshop Globalex a LREC 2018. https://globalex2018.globalex.link/
[7] Sito web ufficiale del CLARIN. https://www.clarin.eu/
[8] Sito web ufficiale di DARIAH. http://stdl.cnr.it/it/dariah

1.2 Risorse lessicali di partenza

La principale fonte dei dati utilizzati per realizzare il Sense Inventory italiano è il repository ILC4CLARIN [1], sviluppato presso l'Istituto di Linguistica Computazionale "A. Zampolli" del Consiglio Nazionale delle Ricerche (ILC-CNR), che ospita il nodo nazionale del Consorzio CLARIN Italia (l'Infrastruttura Comune Italiana per le Risorse e le Tecnologie Linguistiche)[9], parte della federazione Europea di CLARIN-ERIC. I dati utilizzati per il trattamento dei lemmi e dei sensi in italiano sono stati estrapolati da due risorse lessicali ospitate nel repository ILC4CLARIN, rispettivamente i lessici italiani PAROLE-SIMPLE-CLIPS e ItalWordNet.

Un'altra risorsa di riferimento è il database di mapping IWNMAPDB, prodotto di un progetto autonomo sviluppato presso l'ILC-CNR e che contiene il collegamento delle due grandi risorse semantiche lessicali per la lingua italiana di cui sopra (PAROLE-SIMPLE-CLIPS e ItalWordNet).

1.2.1 Parole-Simple-Clips

PAROLE-SIMPLE-CLIPS (PSC) è un lessico a quattro livelli che fornisce informazioni fonologiche, morfologiche, sintattiche e semantiche (Roventini e Ruimy, 2008). È stato sviluppato a partire da progetti di ricerca europei e nazionali fondamentali: il progetto LE-PAROLE per i livelli morfologici e sintattici, il progetto LE-SIMPLE per il modello semantico e il lessico e il progetto italiano CLIPS2 per il livello fonologico e l'estensione della copertura lessicale (Roventini, Ruimy, Marinelli *et al.*, 2007). Il modello teorico di PSC si basa sui risultati dei progetti EuroWordNet (EWN) [10] e ACQUILEX[11] e su una versione rivista della teoria del Lessico Generativo di Pustejovsky (Pustejovsky 1995). È stato codificato a livello semantico in piena conformità con gli standard internazionali stabiliti nel modello PAROLE-SIMPLE e basato su EAGLES, di cui segue modello, linee guida e criteri di annotazione.

Il modello SIMPLE contiene tre tipi di entità formali (Lenci *et al.*, 2000):

- Unità semantiche. I sensi delle parole sono codificati come unità semantiche, o *USem*. Ad ogni *USem* viene assegnato un tipo semantico dall'ontologia, più altri tipi di informazioni

[9] Sito web ufficiale del CLARIN-IT. https://www.clarin-it.it/it
[10] Sito web ufficiale di EuroWordNet (EWN). https://archive.illc.uva.nl//EuroWordNet/
[11] Sito web ufficiale del progetto ACQUILEX. https://www.cl.cam.ac.uk/research/nl/acquilex/acqhome.html

specificate nel modello associato, che contribuiscono alla caratterizzazione del senso della parola.

- Tipo semantico. Ogni tipo comporta informazioni strutturate, organizzate nei quattro ruoli Qualia (Pustejovsky, 1995): formale (fornisce le informazioni che distinguono un'entità all'interno di un insieme più ampio), costitutivo (esprime una varietà di relazioni che riguardano la costituzione interna di un'entità), telico - riguarda la funzione tipica di un ente, cioè a cosa serve l'ente), agentivo (riguarda l'origine di un ente, o il suo venire in essere). Le informazioni Qualia sono suddivise in quelle che definiscono il tipo semantico e altre informazioni aggiuntive che specificano invece ulteriori componenti di un *Usem*. Le prime sono informazioni che definiscono intrinsecamente un tipo semantico così com'è, dunque a un *USem* non può essere assegnato un certo tipo a meno che il suo contenuto semantico non includa le informazioni che definiscono quel tipo.
- Template: una struttura schematica che il lessicografo usa per codificare un dato elemento lessicale. Il template esprime il tipo semantico, più altri tipi di informazioni. I template hanno lo scopo di guidare, armonizzare e facilitare il lavoro lessicografico. Un insieme di modelli principali è stato preparato come default, mentre quelli più specifici possono essere introdotti nella fase di costruzione dei vari lessici, secondo la necessità di codificare particolari concetti in una data lingua.

SIMPLE fornisce un ricco linguaggio espressivo per la rappresentazione dell'informazione semantica e associa ogni tipo dell'ontologia con un gruppo precisato di informazioni che definisce il tipo stesso; il modello associato a un tipo fornisce così una sorta di interpretazione del tipo stesso.

1.2.2 ItalWordNet

Come definita nella pubblicazione *ItalWordNet goes open* (Bartolini, Monachini e Quochi, 2017), ItalWordNet è una risorsa semantica lessicale nata nel 1996 con un approccio *merge and extend* all'interno del progetto EuroWordNet (EWN) (Vossen, 1998), ampliata nell'ambito del progetto nazionale SI-TAL[12] a cura dell'ILC-CNR e successivamente aggiornata e curata fino al 2011 circa. In ItalWordNet (di seguito IWN) le relazioni e i sinonimi sono unicamente specifici

[12] Presentazione del progetto di ricerca per la realizzazione di un Sistema Integrato per il Trattamento Automatico del Linguaggio (SI-TAL). http://www.ilc.cnr.it/it/content/tal

per la lingua, codificati sulla base di precedenti risorse e indipendentemente dal Princeton WordNet[13] (Fellbaum, 1998), a cui invece si ispirava EWN. Come molti altri wordnet, IWN è strutturato intorno a una nozione di sinonimia che comporta l'intercambiabilità di due parole in un determinato contesto, il *synset*. I sinonimi costituiscono quindi gli elementi di base della risorsa e sono formati da sensi di parola sinonimi corrispondenti a parole singole, parole multiple o acronimi appartenenti alla stessa parte del discorso (Bartolini, Monachini e Quochi, 2017).
In IWN sono stati utilizzati per la PoS i tag part-of-speech a grana fine, basati sul tagset ILC/PAROLE dell'ISST Italian Treebank per CoNLL-2007[14] e conformi, come nel caso di PSC, allo standard internazionale EAGLES.

1.2.3 Database di mapping IWN-SIMPLE

Il *linking process* tra PAROLE-SIMPLE-CLIPS (Ruimy et al., 2003) e ItalWordNet (Roventini et al., 2003) è un progetto in linea con il paradigma di costruzione di una nuova generazione di risorse linguistiche, che si prospetta fruttuoso per i benefici reciproci attesi. IWN può infatti fruire di un'informazione sintattica esaustiva e della descrizione dei tipi semantici fornita da PAROLE-SIMPLE-CLIPS; quest'ultimo sarà a sua volta arricchito dalle relazioni tassonomiche, dalla ricca codifica della sinonimia attraverso *synsets*, oltre che dal collegamento a Princeton WordNet, forniti da IWN. Le due risorse, seppure strutturate secondo modelli lessicali diversi, presentano molti aspetti compatibili che sono stati considerati un buon punto di partenza per realizzare il loro collegamento.
Al momento è stato sviluppato il mapping semi-automatico di entità concrete, eventi ed entità astratte. L'ontologia SIMPLE, costituita di ben 157 tipi semantici, permette una strutturazione più fine del lessico rispetto ai 65 concetti principali dell'ontologia IWN, che riflettono solo distinzioni fondamentali. Dunque, l'insieme di tipi semantici di PSC è stata scelto come input per il processo di mappatura, che è orientato come SIMPLE-CLIPS → IWN. La risorsa IWN viene esplorata alla ricerca di candidati per il collegamento che abbiano lo stesso PoS e la cui classificazione ontologica rientri tra le corrispondenze stabilite tra le classi dei lessici (Roventini e Ruimy, 2008). Il mapping tra le due risorse non è completo, e comporta tutte le difficoltà linguistiche e computazionali del caso. Inoltre, nel database sono inseriti anche mapping tra PSC e altre risorse

[13] Sito web ufficiale del Princeton WordNet. https://wordnet.princeton.edu/
[14] The ISST Italian Treebank at CoNLL-2007. Fine-grained part-of-speech tags. http://medialab.di.unipi.it/isst/POS.html

che non sono IWN, che non sono stati però sfruttati in corso d'opera, perciò non verranno trattate in questo elaborato. Quindi, nella presente relazione saranno approfondite solo la modalità di fruizione del database per il Sense Inventory e il task di valutazione di possibili mapping di sensi secondo la similarità di definizione, elaborati durante questo progetto di tesi.

1.3 Perché un Sense Inventory?

La ricerca nell'e-lexicography non è ancora supportata da un'infrastruttura in cui i dati semantici di qualità provenienti dai dizionari possano essere collegati, condivisi, distribuiti e memorizzati su larga scala. Emerge una chiara esigenza di un più ampio e sistematico scambio di competenze, di stabilire standard e soluzioni comuni per lo sviluppo e l'integrazione delle risorse lessicografiche. È necessario ampliare il campo di applicazione di queste risorse di alta qualità a una comunità più ampia, che comprende il Semantic Web, l'intelligenza artificiale, l'NLP (Natural Language Processing) e le scienze umane digitali. ELEXIS mira ad armonizzare questi sforzi e a sviluppare strumenti accessibili, per ridurre il costo e il tempo necessari per aggiornare le risorse esistenti o svilupparne di nuove, a promuovere l'utilizzo di standard di lavoro condivisi e ad aumentare la qualità del proprio operato.

Con il fine di produrre e offrire dati semantici di qualità nell'era digitale in linea con i princìpi di ELEXIS, questo progetto di tesi vuole contribuire all'inclusione di risorse lessicografiche esistenti nella famiglia dei dati open-source affidabili e validi, con l'obiettivo finale di partecipare alla formazione di un corpus multilingue di testi annotati semanticamente per sensi. La creazione di un Sense Inventory si inserisce perfettamente in questo contesto, con la produzione di un inventario sintetico e strutturato di dati semantici e sensi selezionati dai due lessici fondamentali dell'italiano, ItalWordNet e PAROLE-SIMPLE-CLIPS; sarebbe destinato a rendere più lineari e produttivi i task di annotazione automatica per il corpus parallelo, al momento in fase di lavorazione all'interno del progetto ELEXIS. Nello specifico, il Sense Inventory fornirebbe le fondamenta per la codifica semantica della sezione italiana del corpus.

L'altro obiettivo di questo lavoro è sperimentare i nuovi strumenti sviluppati negli ambiti del trattamento automatico del linguaggio (TAL) e della linguistica computazionale, affinché possano essere sfruttati per migliorare l'utilizzo dei set di dati preesistenti ed integrarli con dati nuovi. Per questo, è stato testato un ampliamento del mapping di sensi contenuti nella già citata

risorsa IWNMAPDB. Attraverso algoritmi *word2vec* e l'impostazione di una soglia di similarità tra definizioni, si è cercato di trovare dei sensi candidabili al mapping, per fornire un supporto al complesso compito di controllo e rifinitura del database.

1.4 Sviluppo del progetto

Il piano di lavoro è stato articolato in tre fasi principali: un primo approccio alle risorse disponibili tramite analisi, studio e osservazione, l'elaborazione di risorse create *ex novo* a partire da quelle fornite (il Sense Inventory), infine la valutazione dell'operato e l'investigazione sulle possibilità di miglioramento delle risorse originali.

Inizialmente, le raccolte e i lessici di partenza sono stati osservati e studiati tramite operazioni di stampa e *merge* di dati, formando dei primi set e inventari di sensi strutturati semplicemente ad elenco. In questa prima fase, è stato definito un set di *query* sui database contenenti i dati lessicografici, oltre allo sviluppo di codice per la lettura ed estrazione di informazioni dal dataset annotato. Si annovera qui anche il task di annotazione manuale (effettuato in collaborazione con la collega Irene Pisani), che si è rivelato particolarmente redditizio per la raccolta di informazioni sulla granularità dei sensi contenuti nel dataset e per testare la copertura dei database lessicali. Attraverso quest'ultima indagine manuale delle risorse, si è riusciti ad approfondirne lo studio, per esempio notando alcuni errori nella selezione delle frasi o nel tag di lemmi della versione del dataset generata automaticamente e non ancora corretta.

La seconda fase ha coinciso col cuore del progetto, il trattamento dei dati estrapolati dalle risorse e la creazione del Sense Inventory vero e proprio; il codice e la modalità di svolgimento sono stati modificati e ricalibrati più volte, assecondando le richieste di ELEXIS o le necessità derivate da particolarità o errori riscontrati nell'analisi degli *output*. Lo svolgimento e i risultati ottenuti in questa fase saranno approfonditi all'interno del secondo capitolo.

A conclusione del progetto, è stata condotta una sperimentazione aggiuntiva riguardante il database di mapping, puntando quindi con questo lavoro non solo alla creazione a livello computazionale di risorse fondamentali, ma anche ad un miglioramento e aggiornamento delle esistenti. Il task aggiuntivo consiste nella ricerca di una soglia (*Threshold*) per la valutazione dei possibili nuovi mapping tra lemmi, tra quelli risultati all'interno del Sense Inventory, attraverso l'utilizzo della libreria NLP spaCy. Con questa proposta di ampliamento del mapping di sensi

tramite algoritmi *word2vec*, si è cercato di fornire un supporto al complesso compito di controllo e rifinitura del database di mapping di sensi, proponendo un metodo basato sull'utilizzo di strumenti di Trattamento Automatico del Linguaggio (TAL) che punta al raffinamento delle risorse attraverso la fruizione di strumenti innovativi ed efficienti. Quest'ultimo sviluppo sarà trattato più ampiamente nel capitolo dedicato, il terzo di questo elaborato. Infine, sono state stilate le valutazioni e le statistiche sulla totalità delle risorse lessicografiche ottenute e fruite, riportate nel capitolo finale.

1.5 Introduzione al Word Sense Disambiguation

Il task di Word Sense Disambiguation (WSD) consiste nell'associare le parole in un determinato contesto con il senso più adatto, preso da un inventario di sensi come quello sviluppato in questo progetto di tesi. Può essere visto come un compito di classificazione: i sensi delle parole sono le classi, e un metodo di classificazione automatica è usato per assegnare ogni occorrenza di una parola a una o più classi sulla base di prove dal contesto e da fonti di dati esterne. Il WSD mira a rendere esplicito il significato sottostante alle parole nel contesto in modo computazionale, per eliminare l'ambiguità che di solito non influenza la comprensione umana del linguaggio ma causa molti problemi nell'automatizzazione dei task di NLP. Perciò è comunemente norma che, per permettere una valutazione e un confronto oggettivi dei sistemi WSD, i sensi devono essere enumerati in un inventario di sensi con approccio enumerativo (Navigli, 2009).

All'interno del progetto ELEXIS, questo task ha un alto valore dimostrativo per avvalorare quanto sia forte l'impatto dei dati lessicografici di qualità dell'infrastruttura ELEXIS in compiti chiave di Natural Language Processing (NLP), quali appunto la disambiguazione del senso delle parole, l'Entity Linking e il parsing semantico. Possiamo distinguere due varianti del generico compito WSD (Navigli, 2009):

- *Lexical sample* (a campione lessicale; anche chiamato WSD mirato o *targeted*), dove un sistema deve disambiguare un insieme ristretto di parole target; di solito si trovano una per frase. Solitamente in questa variante sono impiegati sistemi supervisionati, che possono essere addestrati utilizzando un certo numero di informazioni etichettate a mano (*training set*) e poi utilizzati per classificare un set di esempi non etichettati per verificarne le capacità di annotazione (*test set*).

- *All words*, dove ci si aspetta che i sistemi disambiguino tutte le parole di classe aperta in un testo (cioè nomi, verbi, aggettivi e avverbi). Questo compito richiede sistemi ad ampia copertura. Di conseguenza, i sistemi puramente supervisionati possono essere penalizzati per scarsità dei dati, poiché è improbabile che sia disponibile un *training set* che copra l'intero lessico della lingua di interesse.

Come già anticipato, uno dei problemi con cui ci si scontra più spesso nel WSD è la mancanza di risorse adeguatamente ampie e aggiornate; non di rado gli inventari di sensi sono attualmente poco forniti, manchevoli o di lemmi ad alta frequenza o specialistici, con poche o talvolta errate voci nei mapping. Ciò non accade per noncuranza o scorretta gestione delle risorse; lo sforzo necessario per la manutenzione di questi strumenti è ingente e costoso e stare al passo con la richiesta di dati per le nuove tecnologie in fieri è sempre più difficile. Questa situazione non permette di valutare in maniera opportuna gli algoritmi di disambiguazione del WSD e si rischia di ottenere dei risultati falsati e non affidabili. La soluzione più efficiente a questa mancanza è di migliorare, allineare o integrare le risorse già esistenti e di armonizzarle e adattarle ai compiti e ai workframe in cui verranno utilizzate, per ottimizzare e affinare basi già riconosciute e armonizzate in più sistemi.

In questo caso di studio, la creazione di un Sense Inventory a partire da database lessicali ospitati e curati nel repository del CLARIN ed integrati con altri sistemi come Babelnet[15], così come l'esperimento sul possibile ampliamento del mapping tra le due grandi risorse lessicali per l'italiano disponibile nel database dell'ILC-CNR, sono finalizzati al potenziamento dei compiti di Word Sense Disambiguation e alla riorganizzazione del materiale complessivo in italiano a disposizione di ELEXIS. Infatti, tra gli elementi principali di WSD troviamo: la selezione dei sensi delle parole (es, le classi di cui sopra), l'uso di fonti di conoscenza esterne (motivo per cui è fondamentale un Sense Inventory curato e aggiornato), la rappresentazione del contesto e la scelta di un metodo di classificazione automatico.

[15] Sito web ufficiale di Babelnet. https://babelnet.org/

2 La creazione del Sense Inventory

Un Sense Inventory è una risorsa contenente un inventario di sensi associati ai propri lemmi di riferimento di una determinata lingua e ad altri tipi di informazione semantica: classi semantiche generali (come persona, luogo, istituzione ecc.), informazioni di tipo temporale, ecc. (Lenci *et al.*, 2005). Questa tipologia di risorsa ha il fine di fornire dati per l'annotazione semantica di corpora testuali nella lingua data e allo stato attuale può essere formato solo a livello computazionale, dato l'elevato costo di sviluppo dovuto alla necessità di attingere da più lessici per l'estrapolazione dei dati da riorganizzare e rielaborare. Un esempio e pietra miliare è WordNet, il lessico computazionale per l'inglese americano sviluppato a Princeton (USA) e di cui, per l'italiano, TRESSI possiede anche un livello di annotazione semantica, con sensi desunti da ItalWordNet. Tramite WordNet come risorsa di riferimento sono stati annotati anche nomi, verbi e aggettivi di SemCor, una porzione dell'inglese Brown Corpus (ovvero il primo grande corpus strutturato di generi diversi), con il loro senso.

Nello specifico, in questo progetto di tesi è stato sviluppato un inventario di sensi costituito delle informazioni semantiche estratte dai lessici di IWN e PSC, disponibili e accessibili tramite il repository di ILC4CLARIN. Il Sense Inventory in oggetto è basato sui lemmi presenti in un dataset italiano in formato CoNLL-U di proprietà di ELEXIS, che verrà più approfonditamente descritto in seguito, ed è stato creato attraverso diverse fasi di estrazione e trattamento dei dati, con l'aggiunta di un lavoro di integrazione e mapping di sensi come introduzione a un futuro ampliamento della risorsa.

2.1 Il dataset annotato

Il corpus multilingue da annotare è stato costruito dai ricercatori e le ricercatrici che collaborano al progetto ELEXIS; si tratta di un corpus parallelo di frasi estratte dal web (Wikipedia) e tradotte nelle dieci e più lingue delle risorse lessicografiche in costruzione. Dalla sezione del corpus italiana è stato ricavato il dataset di lemmi annotati che funge da base per il Sense Inventory al centro di questa tesi. La risorsa è stata curata dal Dipartimento di Informatica de La Sapienza di Roma guidata dal dott. Roberto Navigli e dai ricercatori e le ricercatrici dell'ILC-CNR di Pisa, tra cui le dottoresse Frontini, Quochi e Monachini che hanno seguito anche lo sviluppo del lavoro qui presentato.

Il dataset su cui si basa il formato finale del corpus è derivato da una revisione manuale della risorsa originale, che invece consisteva in traduzione italiana e tagging automatico di informazioni di un insieme di frasi facenti parte del corpus parallelo di cui sopra. Dalle frasi di cui si compone la risorsa sono dunque stati estratti tutti i lemmi, che sono stati poi taggati seguendo il formato CoNLL-U[16]:

- ID: indice di parola, intero a partire da 1 per ogni nuova frase; può essere un intervallo per i token di più parole; può essere un numero decimale per i nodi vuoti.
- FORM: forma della parola o simbolo di punteggiatura
- LEMMA: lemma o radice della forma della parola.
- UPOS: tag Part-of-Speech, secondo lo standard UD (Universal Dependencies)
- XPOS: tag PoS specifico della lingua; underscore se non disponibile.
- FEATS: elenco di caratteristiche morfologiche dall'inventario universale delle caratteristiche o da un'estensione definita specifica per la lingua; underscore se non disponibile.
- HEAD: head della parola corrente, che è o un valore di ID o zero.
- DEPREL: relazione di dipendenza universale dall'HEAD (radice se la HEAD è 0) o da un sottotipo definito specifico della lingua.
- DEPS: grafo di dipendenza avanzato sotto forma di una lista di coppie HEAD-DEPREL.
- MISC: qualunque altro elemento di annotazione

Nel formato conclusivo in estensione .tsv sono stati modificati manualmente i lemmi e i tag errati che erano stati generati automaticamente, usufruendo anche di un'interfaccia di correzione (fig. 1) sviluppata per questo task nel contesto di Babelscape[17]. Il dataset nella sua forma finale si compone di 2071 frasi, per un totale di 4424 lemmi con diverse PoS. Nella fase successiva di costituzione del corpus annotato semanticamente, un annotatore manuale dovrà selezionare il senso corretto per ogni lemma trovato, scegliendolo dal Sense Inventory appositamente creato a partire da risorse lessicografiche esistenti.

Della copertura del dataset rispetto ai lessici di riferimento per il Sense Inventory si parlerà più ampiamente nell'ultimo capitolo.

[16] Formato standard CoNLL-U. https://universaldependencies.org/format.html
[17] Sito web ufficiale di Babelscape. https://babelscape.com/

fig. 1: interfaccia di correzione con frase di esempio

Il merge di informazioni per il Sense Inventory è stato generato dalla selezione di ogni senso associato alle coppie lemma-PoS nel dataset iniziale, non ancora corretto, in formato CoNLL-U, presi sia dal database *simplelexicon* sia da *newiwn*; a questi poi sono stati aggiunti anche i sensi mappati che si trovano in *iwnmapdb*. Nel momento in cui è stato fornito un formato finale per strutturare il Sense Inventory, si è testato lo stesso programma anche sul dataset definitivo rifinito manualmente dagli specialisti e le specialiste di ELEXIS di cui sopra.

2.2 *Parsing* dei dati iniziali

Il linguaggio di programmazione utilizzato nel progetto è Python 3.8.3 [4]. Per la connessione ai database si è utilizzato il modulo Python di MySQLdb [6]; le righe di inizializzazione e connessione del *cursor* sono identiche in tutti i programmi.

Durante la preparazione dei dati per le query sono sorte due questioni. La prima è la diversa trascrizione delle parole tronche in IWN (con la lettera accentata classica, es. *à*, *è*, *ì*, etc.) e in PSC; in quest'ultimo, infatti, l'accento sull'ultima sillaba è riportato come vocale base con l'aggiunta dell'apostrofo (es. *a'*, *e'*, *i'*, etc.). Ovviamente, questo rende necessaria la trasformazione delle parole tronche prima della query, che si basa proprio sulla corrispondenza tra il lemma fornito in codifica UTF-8 ed estrapolato dal dataset e il lemma contenuto nella tabella del database. Per ovviare a questo problema, si è inserita in tutti i programmi una funzione

di conversione, tramite la quale la variabile contenente la stringa che diverrà parametro della query su PSC viene modificata.

```
def accent(string):
    switch = {('à', "a"+"'"),
              ('è', "e"+"'"),
              ('é', "e"+"'"),
              ('ì', "i"+"'"),
              ('ò', "o"+"'"),
              ('ù', "u"+"'")}
    for case in switch:
        if(case[0] in string):
            s = re.sub(case[0],case[1],string)
            return(str(s))
    return(str(string))
```

codice 1: funzione per la conversione degli accenti

La seconda problematica riguarda la differenza tra frameworks scelti per codificare le PoS nei due database: infatti, né PSC né IWN seguono la codifica di Universal Dependencies[18], che invece è lo standard del progetto ELEXIS per il corpus oggetto di studio e ovviamente per i dataset forniti. Non è conveniente inserire nella query anche la condizione di somiglianza con la PoS; di conseguenza, si è deciso di sfruttare il modulo *re*[19] per le operazioni con Regular Expressions, convertendo le PoS delle tuple ricavate dalla query in UD e confrontandole con la funzione predefinita *fullmatch*, per capire se il senso ottenuto è di nostro interesse (codice 2). In questo modo, la codifica PoS utilizzata in ogni database viene armonizzata con lo standard richiesto e inoltre il lavoro di comparazione dei dati risulta più semplice.

```
def upos(pos):
    switch = {('A', "ADJ"),
              ('AG', "ADJ"),
              ('AV', "ADV"),
              ('N', "NOUN"),
              ('V', "VERB")}
    for case in switch:
        match=re.fullmatch(str(case[0]), pos)
        if(match):
```

[18] Sito web di riferimento per Universal Dependencies. https://universaldependencies.org/
[19] re — Regular expression operations. https://docs.python.org/3/library/re.html

```
            s = re.sub(case[0],case[1],pos)
            return(str(s))
    return(str(pos))
```

codice 2: funzione per la conversione delle PoS in Universal Dependencies

Dopo diversi tentativi, si è infine giunti alla creazione di un programma efficiente per la creazione automatica di un Sense Inventory, adattato poi al formato finale richiesto da ELEXIS in cui per ogni coppia lemma – PoS vengono stampati:

- tutti i sensi di *simplelexicon* non mappati
- tutti i mapping rinvenuti in *iwnmapdb*
- tutti i sensi non mappati di *newiwn*

seguendo questo schema, in cui tutti i campi senza valore vengono riempiti con la dicitura *None* e gli spazi tra i campi corrispondono a *tab* (\t):

```
LEMMA POS   DEFINIZIONE CONCATENATA PSC-IWN    USEMID PSC  DEFINIZIONE     PSC
      ESEMPIO PSC TIPO SEMANTICO PSC      SYNSETID IWN      SENSEID         IWN
      DEFINIZIONE IWN
```

In questo modo, si segue la direzione adottata per la mappatura contenuta in *iwnmapdb* (Roventini e Ruimy, 2008), procedendo dai tipi semantici di PSC, al mapping, ai sensi rimanenti in IWN (synset e iperonimi).

La ricerca e l'estrazione di informazioni ha visto il seguente utilizzo dei database: nell'insieme di tabelle di *simplelexicon* [2], ci si è concentrati sull'estrazione di dati da tre di esse: *usem, usemtemplates* e *templates*. Dall'operazione di left join su *usemtemplates* e *templates* si è ricavato il tipo semantico di ogni senso; tutte le altre informazioni sono state estrapolate da *usem*: l'id, la Part Of Speech (PoS), l'esempio o gli esempi correlati e la definizione. Inoltre, gli *usemid* selezionati da PSC hanno fornito la base per la ricerca via query di mappature in *iwnmapdb*. Nel database *newiwn* [3] è bastato consultare la tabella *wordsxsensesxsynsets*, da cui provengono tutte le informazioni: l'ID del synset a cui appartiene il senso; il *senseid* che disambigua il singolo senso nel synset di cui fa parte; anche qui il lemma, la PoS e ovviamente la definizione del senso. Nel database di mapping *iwnmapdb*, le query sono state svolte tutte su *iwn2psc*, da cui a partire dagli *usemid* scelti abbiamo sono state estratte le mappature e, oltre il suddetto ID dello *USem*, il *synsetid* ad esso associato, sia per la stampa di diverse versioni del Sense Inventory sia per il task finale di mapping. Dal momento che in *wn2psc* sono presenti solo questi dati e la PoS

corrispondente, tutte le informazioni correlate, come ad esempio le definizioni, sono state ricavate da query composte sulle tabelle di IWN e PSC.

2.3 Il formato finale del Sense Inventory

Il programma che costruisce il Sense Inventory seguendo le indicazioni finali del progetto ELEXIS esegue i seguenti task: l'estrazione e ordinamento dei lemmi del dataset corretto, la ricerca di tali lemmi nei database di IWN e PSC, la ricerca di mappature in IWNMAPDB, infine la stampa secondo le linee guida fornite.

Nel primo script per la creazione il Sense Inventory, avendo a disposizione la versione del dataset non ancora corretta manualmente e con estensione .conllu, si era usufruito della libreria pyconll [5] per il parsing del documento e la selezione dei dati interessanti, ovvero il lemma e la PoS corrispondente. Considerando l'estensione .tsv della versione finale del set di dati, invece, si è preferito inserire un parsing *ad hoc*, ottenuto con la funzione predefinita di Python *split* e l'utilizzo di liste per salvare il contenuto estrapolato, per non riconvertire l'estensione del file. Una volta filtrati solo i lemmi con PoS d'interesse (dunque solo aggettivi, avverbi, nomi e verbi), non numerali ed epurati di segni d'interpunzione che causerebbero problemi nelle query sui database, questi sono stati inseriti in un dizionario contenitore che segue la grammatica della libreria *defaultdict* da *collections*[20], il quale è stato successivamente ordinato alfabeticamente sui lemmi contenuti attraverso una funzione molto semplice, spesso utilizzata con alcune varianti per il task di ordinamento di dizionari.

```
def ordina(dict):
    return sorted(dict.items(), key = lambda x: x[1][0])
```

codice 3: funzione per l'ordinamento di un dizionario (*dict*) in Python

Il dizionario così ottenuto, denominato *lemmas*, ha assunto la funzione di vocabolario di riferimento per il programma; dunque, scorrendo tutti gli elementi di *lemmas* con un'iterazione *for*, ogni coppia lemma-PoS è stata cercata nei database IWN e PSC attraverso le operazioni di cui di seguito.

[20] Documentazione del modulo *collections*. https://docs.python.org/3/library/collections.html

Ogni elemento è costituito, come da grammatica, da una coppia *key-value*. Nel caso di *lemmas*, la *key* corrisponde ad un ID numerico crescente; il *value* invece è una lista Python che comprende una coppia lemma-PoS estratta dal dataset e convertita in UTF-8. Le query sono basate sulla ricerca di un lemma all'interno del database che corrisponda a quello su cui attualmente si trova l'iterazione; se c'è una corrispondenza, allora avviene la conversione della PoS acquisita dal database attraverso la funzione *upos* e la comparazione con la PoS estratta da *lemmas*. Prima di poter diventare parametro di una query, il lemma viene però sottoposto a dei controlli per caratteri speciali che spesso precludono un esito positivo nell'output, in particolare per convertire in ASCII quelli che causano errori Unicode, per essere processati correttamente dal programma.

```
if(UnicodeEncodeError):
          lemmaIWN = ud.normalize('NFKD', lemma).encode('ASCII', 'ignore')
```

codice 4: controllo sugli errori Unicode

A questo punto vengono mandate le query, prima su *newiwn* e poi, modificando gli accenti secondo quanto detto sopra con la funzione *accent*, su *simplelexicon*; di ogni tupla risultata dalla query, selezioniamo il campo della PoS e lo confrontiamo con quello fornito inizialmente dal dizionario di lemmi, ovviamente convertendolo in UPoS (UD) come già spiegato all'inizio del capitolo. Usando *fullmatch* si controlla la corrispondenza che se confermata porta al salvataggio della tupla in un dizionario di lemmi di IWN o PSC, che possiedono come *key* rispettivamente il *senseid* o lo *usemid*. Nel caso di definizione mancante, cosa non rara, al suo posto verrà salvata la dicitura *None*. In PSC, se vi è un *match* tra le PoS dev'essere aggiunta una query con left join per selezionare anche il tipo semantico e inserire anch'esso nel *value* del dizionario (codice 5).
Con un altro pezzo di codice di controllo vengono poi selezionati tutti gli *usemid* che non si trovano nella lista *mapdb*, corrispondenti dunque ad un senso non mappato, che vengono stampati secondo le modalità richieste da ELEXIS e già trascritte nella sezione Programmi (codice 6). Con un secondo *check*, di tutti gli *usemid* che invece risultano mappati con un *synset* di IWN vengono selezionati i dati presi dal dizionario di PSC (es. definizione, esempio, tipo semantico, etc.) e salvati in un'altra lista. Tramite i *synsetid* mappati si fa la stessa operazione sul dizionario di IWN e successivamente vengono stampate anche tutte le mappature secondo le modalità standard. Infine, si eliminano dal dizionario di IWN tutti gli *items* che presentano al loro interno un *synsetid* mappato, e quindi già stampato nel documento, così da poter poi stampare

l'intero dizionario senza iterazioni né modifiche. Questo avviene semplicemente tramite l'utilizzo di una lista, *delete*, che raccoglie al suo interno tutti gli elementi di *mapdb* che sono stati già stampati (codice 7).

```
lemmaSL = accent(lemma)
            lemmaSL = lemmaSL.encode("utf-8")
            cursor.execute("SELECT idUsem, naming, pos, exemple,definition
FROM simplelexicon.usem WHERE naming LIKE (%s) ", (lemmaSL, ))
            row2=cursor.fetchall()
            for row in row2:
                if(row is not None):
                    lemma2=row[0]
                    pos2=str(row[2])
                    pos2=upos(pos2)
                    match=re.fullmatch(str(pos2),pos)
                    if(match):
                        cursor.execute("SELECT template FROM
                        simplelexicon.usemtemplates LEFT JOIN
                        simplelexicon.templates ON
                        usemtemplates.idTemplate=templates.idTemplate WHERE
                        usemtemplates.idUsem LIKE (%s)  ", (lemma2, ))

                        template=cursor.fetchone()

                        if(template is not None):
                            templ=template[0]
                        else:
                            templ = 'None'
                        if(row[4] is not None):
                            psc[lemma2]=[row[1], pos2, row[3], row[4],
                              templ]
                        else:
                            psc[lemma2]=[row[1], pos2, row[3], 'None',
                              templ]
```

codice 5: query su *simplelexicon* (parte del procedimento è corrispondente a quello per *newiwn*)

```
for elem in psc.items():
```

```
                usemid = elem[0]
                usemid = usemid.encode("utf-8")
                cursor.execute("SELECT      synset1id,      word2id      FROM
iwnmapdb.iwn2psc WHERE word2id LIKE (%s) ", (usemid,))
                row3=cursor.fetchall()
                if(row3):
                    for row in row3:
                        if(row is not None):
                            mapdb.append(row)
```

codice 6: query sul database di mapping a partire dai lemmi presenti in PSC

```
if(delete):
          for item in delete:
                    del(iwn[item])
```

codice 7: eliminazione dei sensi già mappati dal dizionario di IWN

Nel caso particolare di sensi mappati con definizione perfettamente corrispondente, viene stampata una sola definizione nel campo DEFINIZIONE CONCATENATA PSC-IWN, sempre attraverso un controllo con il modulo *re*.
Parte della risorsa del Sense Inventory è consultabile nell'appendice di questo elaborato.

2.4 Considerazioni sulla performance del Sense Inventory

In conclusione, il numero totale di lemmi presenti nel Sense Inventory, quindi con una corrispondenza tra il dataset CoNLL-U e i lessici di riferimento, ammonta a 3860, su 4424 lemmi del corpus. I sensi e i mapping riportati nel Sense Inventory sono 12.944, su un totale di 15.672 sensi estrapolati da PSC e IWN; da IWNMAPDB, invece, sono stati estratti e inseriti nel Sense Inventory come sensi rilevanti 3461 mapping. Di tutti i sensi integrati nell'inventario, 3519 sono legati a verbi, 7142 a nomi, 2137 ad aggettivi e solo 146 ad avverbi.
L'algoritmo di strutturazione del Sense Inventory, allo scopo di evitare duplicazioni nel caso di sensi con mapping, stampa in output direttamente la mappatura e non il singolo senso estratto da uno dei due lessici IWN o PSC; in questo modo, unisce molti sensi tra loro e il numero di risultati nell'output diminuisce di conseguenza.

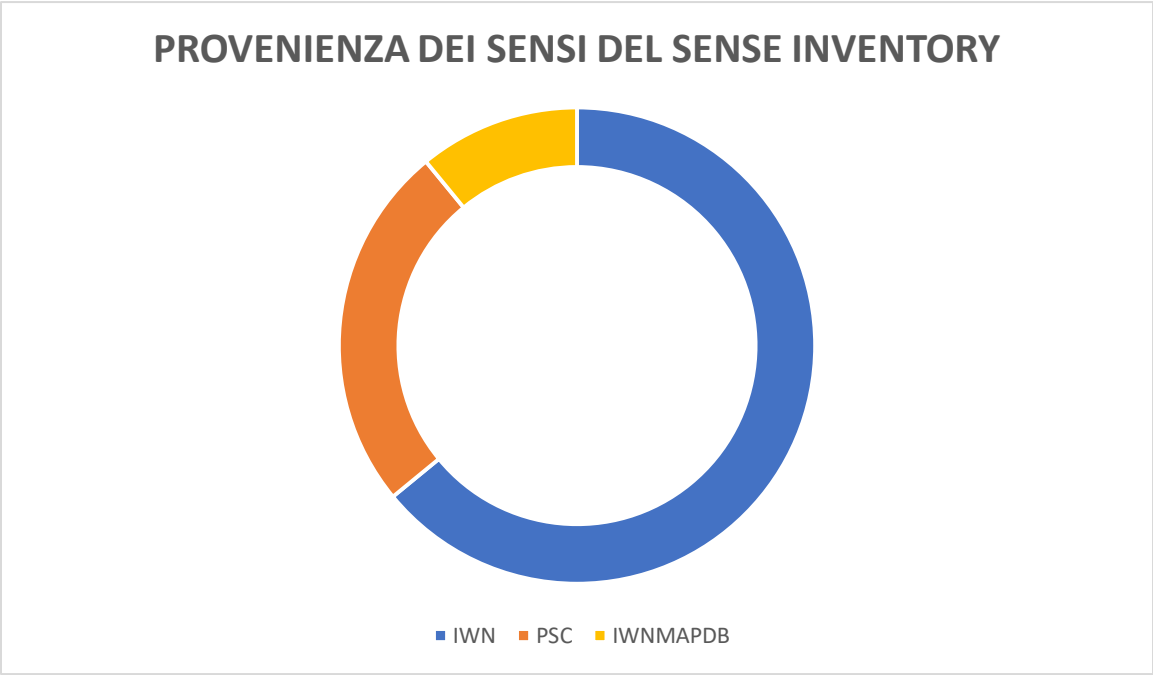

fig. 2: grafico ad anello che mostra la provenienza dei sensi del Sense Inventory

Il file definitivo del Sense Inventory è stato poi ulteriormente valutato: prendendo alcuni gruppi di circa 20 lemmi a campione dall'inventario di sensi, si è effettuato il controllo manuale di tutti i sensi del lemma nei database e si è confrontato il risultato con l'output finale. Attraverso questa operazione è stato possibile scoprire alcune imprecisioni nell'algoritmo di estrazione, come per esempio un erroneo trattamento delle parole accentate in *newiwn*, che ha reso necessaria l'inserzione di un controllo e di una conversione in Unicode ASCII di alcuni caratteri. Dall'ispezione dei documenti in input e output sono emerse anche alcune lacune nel Sense Inventory, attribuibili sia alla presenza di lemmi particolari nel dataset sia a mancanze all'interno dei database. Innanzitutto, sono presenti molte parole inglesi o in altre lingue straniere, ovviamente non pertinenti a lessici di lingua italiana: questo dato acquisisce particolare rilevanza poiché influenza i calcoli statistici sulla copertura dei database rispetto al dataset, che sarà trattata nel quarto capitolo di questo elaborato dove verrà riportata un'analisi valutativa più approfondita delle risorse. Queste osservazioni portano alla conclusione di dover tenere in considerazione una percentuale di errore per lemmi che non sarebbero comunque integrati in alcun lessico di riferimento per un progetto sulla lingua italiana, in questo specifico caso perché non coerenti con la struttura originaria dei database.

Un’altra singolarità è la presenza nel dataset di lemmi contenenti caratteri particolari difficilmente codificabili e leggibili per i programmi Python, sia parole straniere non inglesi come *sentō′* e *ōkimi*, sia numerali come *17º*: questi ultimi sono stati esclusi a priori dagli algoritmi, mentre le parole straniere rientrano nel margine di errore di lemmi sicuramente non coperti nei database.

Inoltre, è particolare la presenza nei lessici e nel dataset CoNLL-U di molti lemmi tecnici o rari, anche se talvolta trattati erroneamente nel tagging automatico del dataset; un esempio sono i nomi di composti chimici poco conosciuti ai non esperti del settore come il *bismuto*, o termini tecnici di scarso utilizzo quotidiano come *acufene*. Invece, non di rado sono assenti nei database alcuni dei sensi più comuni di lemmi ad alta frequenza, o gli stessi lemmi presentano una sola delle PoS possibili mancando di un’altra molto utilizzata; solitamente, in quest’ultimo caso viene a mancare la PoS ADJ o NOUN su un lemma che può essere utilizzato in entrambi i ruoli. Un esempio è la parola *giallo*, che nonostante possa essere il nome comune di un colore (“il giallo è il mio colore preferito”) e venga usato spesso in questa forma, è stato trovato nel dataset e dunque inserito nel Sense Inventory con la sola funzione di aggettivo; o anche *giusto*, che compare nel dataset come aggettivo e anche avverbio, ma mai come nome.

Nel capitolo 4 verrà fornita un’analisi più puntuale dello stato dell’arte delle risorse lessicali di partenza da cui sono stati estratti i sensi, per verificarne la copertura rispetto al dataset e al Sense Inventory.

3 Estensione automatica del mapping di sensi

La strategia utilizzata dall'ILC-CNR per la strutturazione del database IWNMAPDB si basa sulla mappatura semantica semiautomatica tra i lemmi di PSC e IWN guidata dal tipo semantico codificato in PSC, attraverso l'utilizzo del software LINKPSC_IWN. Una caratteristica di questo mapping è che coinvolge elementi lessicali che hanno uno status diverso, cioè unità semantiche (*USem*) e set di sinonimi (*synset*). Il processo di mappatura ha previsto i seguenti passi (Roventini, Ruimy, Marinelli *et al.*, 2007):

- Selezione di un tipo semantico di PSC e definizione dei criteri di scelta, cioè o tutte le sue USem o solo quelle che recano una determinata informazione (ad esempio, una certa PoS o un'ambiguità semantica che mette alla prova il software e che verrà controllata poi manualmente nei risultati).
- Selezione di uno o più vincoli di mappatura sulla base delle corrispondenze stabilite tra le classi concettuali delle due ontologie, al fine di restringere la mappatura automatica.
- Validazione manuale della mappatura automatica e salvataggio dei risultati.
- Se necessario, rilassamento/*tuning* dei vincoli di mappatura e nuova elaborazione dei dati di input.

In questa fase di lavoro di tesi si è testata una metodologia di ampliamento del collegamento e della mappatura tra i sensi di IWN e PSC, attraverso il calcolo automatico di similarità tra le definizioni dei sensi. Questa tecnica mira a integrare i risultati ottenuti con il metodo sviluppato all'ILC-CNR. L'approccio automatico al collegamento tra sensi simili è stato applicato utilizzando la libreria per NLP tasks spaCy [8], in particolare la funzione per vettori di parole e somiglianza semantica *similarity* e il pacchetto *it_core_news_md* [21], all'interno di uno script Python.

```
nlp = spacy.load("it_core_news_md")
[...]
doc1 = nlp(definizione_psc)
doc2 = nlp(definizione_iwn)
simil = doc1.similarity(doc2)
```

codice 8: pseudocodice dell'utilizzo di spaCy all'interno dello script

[21] spaCy: Word vectors and semantic similarity. https://spacy.io/usage/linguistic-features#vectors-similarity

Il programma è suddiviso in due sezioni che svolgono task diversi: la prima analizza diverse mappature già presenti in IWNMAPDB attraverso algoritmi di similarità basati sui *word embeddings* e ne ricava una soglia di similarità (Threshold) da utilizzarsi per i nuovi mapping; la seconda parte testa diversi lemmi senza alcun senso mappato per trovare dei candidati all'integrazione del database di mapping.

Lo scopo di questo compito è di usufruire dei servizi per il Natural Language Processing basati su algoritmi come *word2vec*[22] per testare un diverso metodo di mappatura basata sulla vicinanza semantica fra le descrizioni di sensi, per integrare alcuni lemmi non mappati testando una strategia di mapping diversa, più semplice ma che permette di velocizzare il collegamento tra unità semantiche o sensi con definizioni estremamente simili nella forma.

I lemmi passati come input per il programma non sono stati selezionati casualmente da un algoritmo *random*, ma scelti manualmente per garantire un campionario diversificato delle diverse casistiche: lemmi con pochi o molti mapping tra i sensi correlati nelle risorse, che presentino definizioni identiche o estremamente diverse. I candidati sono stati selezionati da un elenco di lemmi con almeno un senso presente nel database di mapping, generato automaticamente a partire dal Sense Inventory con un algoritmo.

3.1 Soglia con lemmi mappati

La prima parte dell'algoritmo per il task di mapping si basa sull'utilizzo della funzione *similarity* della libreria spaCy [8] per l'osservazione di mappature già estratte dal database IWNMAPDB. Lo scopo finale è di calcolare una soglia per i sensi non mappati, ovvero un valore-limite che garantisca una percentuale di similarità tra le definizioni tale da poter considerare i due sensi analizzati come assimilabili tra loro, così da aggiungere candidati alla mappatura.

I lemmi importati come input del programma sono stati selezionati manualmente da una lista di coppie lemma - PoS con almeno un senso mappato, quest'ultimo estratto dallo script per le statistiche finali di cui si parlerà nell'ultimo capitolo. La scelta dei lemmi non è stata affidata a funzioni di selezione *random* da dati perché si è preferito ottenere un elenco vario, che prendesse in considerazione diverse tipologie di mappature: alcuni sensi presentano una perfetta corrispondenza delle definizioni, tanto che nella voce *DEFINIZIONE CONCATENATA PSC-IWN*

[22] Google Code Archive, *word2vec: tool for computing continuous distributed representations of words.* https://code.google.com/archive/p/word2vec/

del Sense Inventory ne viene riportata una sola; altri casi sono la presenza di un sintagma identico all'interno di una frase più articolata, di alcune parole in comune, o anche di soli sinonimi senza effettive corrispondenze. La variazione nelle casistiche è stata creata appositamente per testare gli algoritmi con *word2vec* di spaCy, il quale assegna a ogni token un *embedding* pre-addestrato su grandi corpora di *training*; sulla base dei vettori di parole così ottenuta viene calcolata la similarità semantica tra le frasi.

Word2vec è una semplice rete neurale artificiale a due strati progettata e utilizzata nell'ambito NLP, basata a sua volta sulle architetture CBOW e Skip-Gram. L'algoritmo prende in input un corpus e restituisce un insieme di vettori, i *word embeddings*, estraendo informazioni di semantica distribuzionale e inviando tali informazioni ad una rete neurale ricorrente supervisionata. La dimensione del *word embedding* è determinata dal numero di nodi nell'*hidden layer* della rete neurale. A dimensioni maggiori corrispondono rappresentazioni più dettagliate. I *word embeddings* possono essere poi usati come codifiche delle parole da inviare in ingresso ad un sistema di *machine learning* (per es. un'altra rete neurale). Quella dei *word embeddings* è una rappresentazione distribuita, la cui dimensione (il numero di nodi dello strato nascosto) può essere fissata in funzione dello scopo finale, ed è usabile da modelli terzi [11].

A livello algoritmico, lo script per la soglia è così strutturato: inizializzato una lista di elementi composti (le coppie lemma – PoS d'interesse), viene aperto come input il file con l'elenco di sensi mappati di cui sopra e si raccolgono tutti i *synsetid* e *usemid* il cui lemma e la cui PoS si trovino nella lista iniziale. Aperto poi anche il file del Sense Inventory, ogni riga viene scomposta attraverso la funzione predefinita *split* in una lista, basando la separazione sull'elemento tab (*t*); dopodiché, si confrontano gli elementi della prima lista di lemmi scelti (*testing*) con quelli di ogni riga. Se il lemma, la PoS e gli ID coincidono, le definizioni (prese dalla riga del Sense Inventory) vengono salvate in una variabile, corrette con il punto fermo di fine frase se non già presente, e infine confrontate attraverso la *similarity* (codice 9). Tutti i valori così ottenuti vengono salvati in una lista (*allsimilar*), da cui viene poi ricavato il Threshold calcolando la media aritmetica di tutti i valori di *similarity* in output: arrotondando, si arriva a 0,746. Considerando sia i risultati dei test sui lemmi non mappati (di cui si parlerà nel capitolo successivo) sia uno scarto di errore minimo nei risultati di spaCy, sarebbe ideale settare il threshold per mapping o puntando al minimo di 0,75 stesso o a un valore leggermente più alto, come 0,78.

```
For elem in testing:
            #check di lemma e pos
```

```
if(line[0]==elem[0] and line[1]==elem[1]):
    #check di usemid e synsetid
    if(line[7]==elem[2] and line[3]==elem[3]):
        print("\n", file=out)
        print(line[0], line[1], file=out)
        #aggiungi il punto alle definizioni per migliore risultato
             di similarity
        if(line[4].endswith('.')):
            defpsc=str(line[4])
        else:
            defpsc=str(line[4])+"."

        If(line[9].endswith('.')):
            defiwn=str(line[9])
        else:
            defiwn=str(line[9])+"."
        Doc1=nlp(defpsc)
        doc2=nlp(defiwn)
        simil = doc1.similarity(doc2)
        allsimilar.append(simil)
```

codice 9: selezione delle frasi e confronto tramite spaCy (nlp) per calcolo di *similarity*

3.2 Allineamento di lemmi privi di mapping

Nella seconda parte dello script è stato implementato un test per i lemmi ancora non mappati, sempre basato sulla *similarity* delle definizioni dei sensi. Anche in questa fase, gli input sono stati selezionati manualmente da un elenco generato all'interno del programma per le statistiche finali, attraverso la collezione automatica di tutti i lemmi non mappati in una lista Python.

Considerando la tipologia di task, in realtà sarebbe possibile selezionare i sensi non mappati anche di lemmi che presentano almeno un mapping di un altro senso, situazione che il programma tiene in considerazione inserendo un controllo con query su *iwnmapdb* per evitare di selezionare erroneamente sensi già mappati. In particolare, questo controllo vale per gli *usemid* di PSC, poiché la relazione tra sensi su cui si fonda il database, la quale è di tipo PSC → IWN,

permette che un *synsetid* (un insieme di sinonimi di IWN) possa essere mappato più volte con diverse unità semantiche di PSC, ma ogni *usemid* è solitamente mappato una sola volta con un solo *synset*. Ovviamente sono presenti anche delle eccezioni, come nel lemma *gara* dove la stessa *USem* è mappata con due *synset* diversi.

Alla fine si è deciso di passare in input al programma solamente i lemmi che non presentino alcun senso mappato; è stata fatta questa scelta sia al fine di rendere il programma più efficiente, evitando controlli che poi andrebbero a scartare però buona parte dei dati in elaborazione, sia per evitare di lavorare su sensi che, anche dopo una rivalutazione, probabilmente non sarebbero comunque da considerare come candidati al mapping.

L'algoritmo di *testing* lavora nel seguente modo: l'insieme di coppie lemma – PoS d'interesse è salvata manualmente in una lista Python, con lo stesso metodo utilizzato per il test sui lemmi mappati. Anche in questo secondo task, i lemmi all'interno della lista necessitano di un controllo per verificare che effettivamente non siano mai stati mappati. Per fare questo, si apre in input il file del Sense Inventory e, come già fatto per la ricerca del Threshold, si va a creare una lista per ogni riga di testo; dopodiché, si restringe il campo di ricerca alle righe di sensi o appartenenti solo a PSC o mappati con un controllo sul terzo elemento della riga (che corrisponde allo *usemid*) e si esegue una query su *iwnmapdb* per controllare che non sia mappato.

```
for line in senseinventory:
        line=re.sub(r"\n","", line)
        line=line.split("\t")
        for elem in tocheck:
            nomapids=[]
            if(line[0]==elem[0] and line[1]==elem[1]):
                if(line[3] != 'None'):
                    usem=str(line[3])
                    cursor.execute("SELECT synset1id, word2id FROM
                        iwnmapdb.iwn2psc WHERE word2id LIKE (%s) ",(usem,))
                    row3=cursor.fetchall()
                    if(not row3):
                        nomapids.append([line[0], line[1], usem, line[4]])
```

codice 10: controllo sui lemmi non mappati

Gli *usemid* non mappati vengono salvati in una lista definitiva da cui partono i test sui sensi: per ogni ID, si seleziona la definizione corrispondente sempre sulla riga del Sense Inventory a cui viene all'occorrenza aggiunto il punto fermo. Con una seconda query, si fa una ricerca di tutti i *synset* con gli stessi valori lemma – PoS e si fa un controllo incrociato delle definizioni dello *usemid* con tutti i *synset* scovati, sempre attraverso la funzione *similarity*.

3.3 Risultati del test di mapping

Il modello di semantica distribuzionale su cui si basa *word2vec* nasce dall'idea che lemmi simili si trovino in contesti simili. Il calcolo della similarità tra parole viene operato secondo il principio simmetrico di vicinanza del tipo di relazione semantica che intercorre tra le parole del contesto; dunque, il calcolo di *similarity* può risultare difettoso in casi di relazioni non simmetriche o proprio gerarchiche, per esempio con iponimia o iperonimia. Allo stesso modo saranno molti gli ostacoli per l'algoritmo, quali ad esempio frasi non concluse o semanticamente poco chiare; oppure parole che seppur appartenenti alla stessa famiglia e chiaramente vicine, se non quasi sinonime, per un parlante nativo della lingua, non vengono riconosciute come tali dalla rete neurale.

Da alcuni test sulle frasi del dataset ELEXIS, per esempio, è stato rilevato che la *similarity* aumenta considerevolmente con frasi concluse anche a livello di punteggiatura, dunque nel contesto dell'elaborazione delle informazioni sui lemmi nel codice Python è stato aggiunto un punto fermo alla fine di ogni definizione che non lo possedesse originariamente. Inoltre, è stato notato che l'uso della stessa parola, ma con numero diverso singolare/plurale (es. *America* e *Americhe*), per quanto ovviamente molto simile a livello di forma, interferisce molto nel calcolo della *similarity* in modo negativo, sicuramente a causa o di mancanze nel *training* dell'algoritmo o a causa dei problemi legati al calcolo distribuzionale della semantica di cui sopra.

Durante una revisione manuale delle risorse è inoltre emerso che nessun aggettivo né avverbio è stato ancora mappato. Il 51.49% dei lemmi non mappati con alcun senso, infatti, sono aggettivi o avverbi. Considerando la metodologia adottata finora dall'ILC-CNR (Roventini, Ruimy, Marinelli *et al.*, 2007), bisogna ricordare che con alcuni tipi semantici di PSC non è ancora stato affrontato il task di mapping a causa della complessità del compito, e tra i mancanti sicuramente possiamo annoverare quelli legati ad aggettivi e avverbi. Infatti, dal momento che il mapping su

cui è stata costruita la risorsa lessicale si basa sulla corrispondenza tra tipo semantico e concetto o *synset* di ogni lemma (Roventini e Ruimy, 2008), si cade in una serie di casistiche abbastanza complesse che non permettono una mappatura automatica per PoS come *ADJ* o *ADV*, se non con un grande sforzo manuale, che quindi renderebbe l'aiuto fornito della programmazione quasi vano. Questo accade perché ovviamente nel caso di queste tipologie di lemmi si troverebbero concetti e tipi semantici tutti estremamente simili tra loro, che indicano lo stesso uso semantico del lemma e non si differenziano se non nell'oggetto modificato appunto dall'avverbio o dall'aggettivo, sostanzialmente nel contesto in cui questi siano utilizzabili. La differenza non è segnalabile a livello di informazione semantica primitiva e dunque blocca il processo di mappatura.

Uno scenario futuro potrebbe prevedere un nuovo metodo di mappatura che sfrutti le capacità degli strumenti basati sul *word embedding* e sul calcolo della vicinanza tra vettori di parole come spaCy. Si potrebbe scegliere così di non basarsi sulla correlazione tra i tipi semantici degli *USem* e i concetti dei *synset* correlati allo stesso lemma, ma piuttosto sulla interscambiabilità tra due parole, anche sinonime, all'interno di una frase, valutata sulla base della *similarity* delle loro definizioni. In questo modo, non si andrebbe a creare una mappatura tra sensi perfettamente corrispondenti, ma piuttosto tra sensi tanto vicini da poter essere utilizzati esattamente negli stessi contesti. È però necessario che i sensi valutati siano sinonimi in senso "orizzontale", che non si trovino dunque in un rapporto gerarchico di iponimia o iperonimia, poiché gli algoritmi *word2vec* presentano diverse problematiche con questo tipo di relazione semantica, mentre funzionano meglio nella valutazione di sensi che si trovino esattamente sullo stesso piano.

Un esempio tratto dal *test set* annotato manualmente è il lemma *elevato* (ADJ), dalla seguente frase del dataset: "Il portale contiene una gamma molto ampia di dati aperti di elevato valore relativi ai vari settori d'intervento dell'UE, tra i quali figurano l'economia, l'occupazione, le scienze, l'ambiente e l'istruzione". In questo contesto, il lemma è stato taggato col senso definito come "notevole per qualità" (*synsetid*: 50061; *senseid*: 66784), proveniente dal lessico IWN; ma la stessa coppia lemma – PoS nel lessico PSC non è correlata ad alcun senso che sia adatto al contesto. Infatti, sono presenti solo i sensi aventi come definizione: "alto" (con tipo semantico *Physical Property*, esempio fornito: "montagna elevata"), "eletti, nobili" o "abbiente, facoltoso". In questo caso, si potrebbe per esempio creare un mapping con un senso corrispondente, ma

appartenente ad un altro lemma, come *notevole* nel senso con definizione “degno di essere notato, in particolare per l'eccellenza raggiunta” (*usemid*: USemD6385notevole).

4 Valutazione delle risorse

In questo capitolo saranno analizzate le risorse lessicali utilizzate per la creazione del Sense Inventory, ovvero i database e il dataset, al fine di valutarle nella loro adeguatezza ai compiti che sono preposte a svolgere e a mettere a confronto la quantità e qualità dei dati contenuti in ognuna di esse.

Nella valutazione sono stati fondamentali sia il task di annotazione manuale legato alla creazione di un *test set* di frasi estratte dalla risorsa CoNLL-U, sia la scrittura di un programma per il calcolo di statistiche sulla copertura dei database. In particolare, la ricerca manuale sui database per l'annotazione di frasi dal dataset si è rivelata fondamentale sia per l'esplorazione della struttura dei database stessi sia per comprenderne meglio le mancanze o le potenzialità, così come per scrivere codice che riuscisse a prevedere e gestire la maggior parte degli scenari possibili.

4.1 Annotazione semantica manuale del dataset

Tra i task sviluppati in questa tesi, l'annotazione semantica manuale di alcune frasi selezionate dal dataset ConLL-U ha permesso di approfondire lo studio e l'analisi delle risorse di partenza, e di verificare la copertura dei database e del dataset con cui si è lavorato rispetto alle risorse create, come il Sense Inventory.

Il compito ha previsto la selezione di 50 frasi dal dataset e l'annotazione manuale dei singoli lemmi (nomi, verbi, aggettivi e avverbi) con il senso più adatto al contesto tra quelli estratti da IWN e PSC. Il lavoro di annotazione, che per sua natura prevede due annotatori, è stato svolto in collaborazione con la collega Irene Pisani. Sono state selezionate 25 frasi da una metà del corpus, scelte in base alla copertura del database rispetto ai lemmi di cui sono composte. Ogni annotatrice ha annotato il proprio test-set ed ha poi svolto il task di annotazione anche sulle frasi dell'altra, affinché si potesse calcolare l'*inter annotator agreement*.

Il tagging dei lemmi è avvenuto cercando manualmente sui database ogni lemma con la PoS corrispettiva (già correlata automaticamente nel dataset) e associandovi, se presente nei database, un senso da IWN e da PSC attraverso lo *usemid* o la coppia *synsetid – senseid*. L'annotazione è stata elaborata su un file di estensione .xslx in cui sono riportate le frasi e, in ogni colonna, i dati associati ad ogni lemma. Dunque, la struttura del documento è la seguente:

```
# sent_id = (id frase)
```

```
# text = (frase)
ID    FORM  LEMMA UPOS  XPOS  FEATS HEAD  DEPREL      DEPS  MISC  ELEXIS:PSC
      ELEXIS:IWN
```

Ogni senso mancante viene indicato con la dicitura *missing*. Un senso è mancante o quando cercando il lemma con la propria PoS correlata la query non riporta alcun risultato, o quando tra i sensi presenti nei database non si trova quello adatto al contesto. Spesso, sono anche salvati nei database alcuni sensi che non possiedono alcuna descrizione (valore *null*): in questi casi, sono state adottate come soluzioni una ricerca tra gli iperonimi e i *synset* (nel caso di sensi in IWN), per trovare il significato più vicino e plausibile e scegliere il senso con semantica del *synset* più adatta al contesto, o nel caso di mapping una query sul database *iwnmapdb* per guardare la definizione corrispondente nell'altra risorsa lessicale.

Un caso particolare è quello delle *multiword*, ovvero gruppi di parole corrispondenti a locuzioni ed espressioni polirematiche (o multilessicali). Presenti in entrambi i lessici, ma in prevalenza in IWN, le unità complesse sono risultate gestibili nei modi più diversi, per esempio seguendo le indicazioni del LAW-MWE-CxG 2018 (COLING)[23]. Si è infine stabilito di seguire uno schema elaborato appositamente per rendere il task più agevole, soprattutto per il parsing che il documento avrebbe subìto successivamente: ogni caso di *multiword* in una frase presenta il flag MW prima della trascrizione degli ID del senso di riferimento. Al flag è correlato un numero, che funge da indice della *multiword* nella frase; dunque, se in una frase vi sono più unità multilessicali che si possono trovare nei lessici come un'unica unità semantica, la prima nell'ordine della frase verrà indicata con MW: 0, la seconda con MW:1, e così via. Ogni lemma appartenente alla stessa *multiword* presenta lo stesso flag, con lo stesso numero e ovviamente gli stessi id. Di seguito è riportato l'esempio del lessema complesso "essere umano", presente sia in *newiwn* che in *simplelexicon*:

```
MW: 0 usemid: USemD6498essere_umano | MW: 0 synsetid: 863;  senseid: 1956
MW: 0 usemid: USemD6498essere_umano | MW: 0 synsetid: 863;  senseid: 1956
```

Nel caso in cui una *multiword* non sia presente come tale nelle risorse lessicali, si è provvisto a ricreare il significato attraverso l'attenta selezione di sensi di ogni singolo lemma; se il significato

[23] CUPT format specification. http://multiword.sourceforge.net/PHITE.php?sitesig=CONF&page=CONF_04_LAW-MWE-CxG_2018___lb__COLING__rb__&subpage=CONF_45_Format_specification

della *multiword* non è ricreabile in questa maniera, magari perché si tratta di un senso figurato del lessema, talvolta si è anche scelto di segnare quel lemma con *missing*.

4.1.1 Algoritmo di selezione delle frasi per il *test set*

Per ottimizzare la selezione delle frasi è stato scritto un programma Python che prende in input il dataset in formato CoNLL-U e assegna ad ogni frase il numero di lemmi presenti in entrambi i database di riferimento. L'importanza di scegliere una frase con il maggior numero possibile di lemmi coperti totalmente nei database sta nella maggiore accuratezza del task di annotazione; se vi fossero molti lemmi, ma con sensi presenti in un solo database, l'annotazione presenterebbe molti più valori *missing* e sarebbe meno ricca ed accurata. Inoltre, non sarebbe stato possibile utilizzare il database di mapping per controlli su molti lemmi senza definizione, aumentando i campi *missing* che rendono più difficile la valutazione delle risorse.

L'algoritmo dello script è abbastanza semplice: attraverso la libreria *pyconll*, dal file in input vengono analizzate solamente le frasi con ID superiore a 1037, così da considerare le *sentences* nella metà del set di competenza. Per ogni frase viene inizializzata una variabile contatore che tenga conto di quanti lemmi all'interno della frase possiedano almeno un senso sia in *newiwn* che in *simplelexicon*; dunque, di ogni lemma con PoS d'interesse vengono effettuate delle query sui database e come già spiegato nei capitoli precedenti per altri script anche un controllo sulla PoS; anche in questo programma infatti è presente la funzione per conversione di PoS *upos*. Per collezionare le tuple risultanti dalle query è utilizzata la funzione predefinita *fetchone* anziché *fetchall* come nel programma per la struttura del Sense Inventory poiché non è necessario possedere tutte le tuple corrispondenti al lemma, ma solo una conferma della presenza di almeno un senso in ogni base di dati controllata. Nella chiusura del programma, ogni frase viene inserita col suo conteggio di lemmi totalmente coperti dai database come valore di un dizionario, qui chiamato *frasi*, la cui *key* è l'ID originale della frase così come salvato dal metadato/commento del dataset CoNLL-U *sent_id*.

```
corpus = pyconll.load_from_file("dataset2000wiki_UDPIPE.conllu")
frasi={}
for sentence in corpus:
      counting=0
      sentid=int(sentence.meta_value('sent_id'))
      if(sentid>1036):
```

```
for token in sentence:
    if (token.upos == 'VERB' or token.upos == 'NOUN' or
      token.upos == 'ADV' or token.upos == 'ADJ'):
            word=(token.lemma).strip().encode("utf-8")
            firstpos=str(token.upos)
            cursor.execute("SELECT lemma, pos FROM
            newiwn.wordsxsensesxsynsets WHERE lemma LIKE (%s)",
            (word,))
            catch=cursor.fetchone()
            if(catch):
                  pos=str(catch[1])
                  pos=upos(pos)
                  check=re.fullmatch(pos, firstpos)
                  if(check):
                        cursor.execute("SELECT naming, pos FROM
                        simplelexicon.usem WHERE naming LIKE
                        (%s)", (word,))
                        catch2=cursor.fetchone()
                      if(catch2):
                            pos2=str(catch2[1])
                            pos2=upos(pos2)
                            check2=re.fullmatch(pos2,firstpos)if(
                            check2):
                                  counting=counting+1
frasi[sentid]=[(sentence.meta_value('text')), (counting)]
```

codice 11: attribuzione ad ogni frase del numero di lemmi presenti in entrambi i database

4.2 Analisi comparativa della copertura delle risorse

Il Sense Inventory funge anche da mezzo per un'analisi dell'attuale ampiezza dei database di cui abbiamo usufruito, in particolare per studiare la loro effettiva copertura ed efficienza rispetto alle necessità di ELEXIS e il loro stato dell'arte. In questo capitolo sono presentati dei dati ricavati

dal confronto tra i lessici, il dataset formato CoNLL-U e il Sense Inventory, raccolti ed elaborati con l'aiuto di un programma Python.

Dall'output dello script sappiamo che i lemmi contenuti nel dataset definitivo sono 4424 tra aggettivi (1000), avverbi (231), nomi comuni (2418) e verbi (775), di cui 3860 sono stati poi inseriti nel Sense Inventory per corrispondenza nei lessici come già accennato nelle considerazioni alla fine del secondo capitolo.

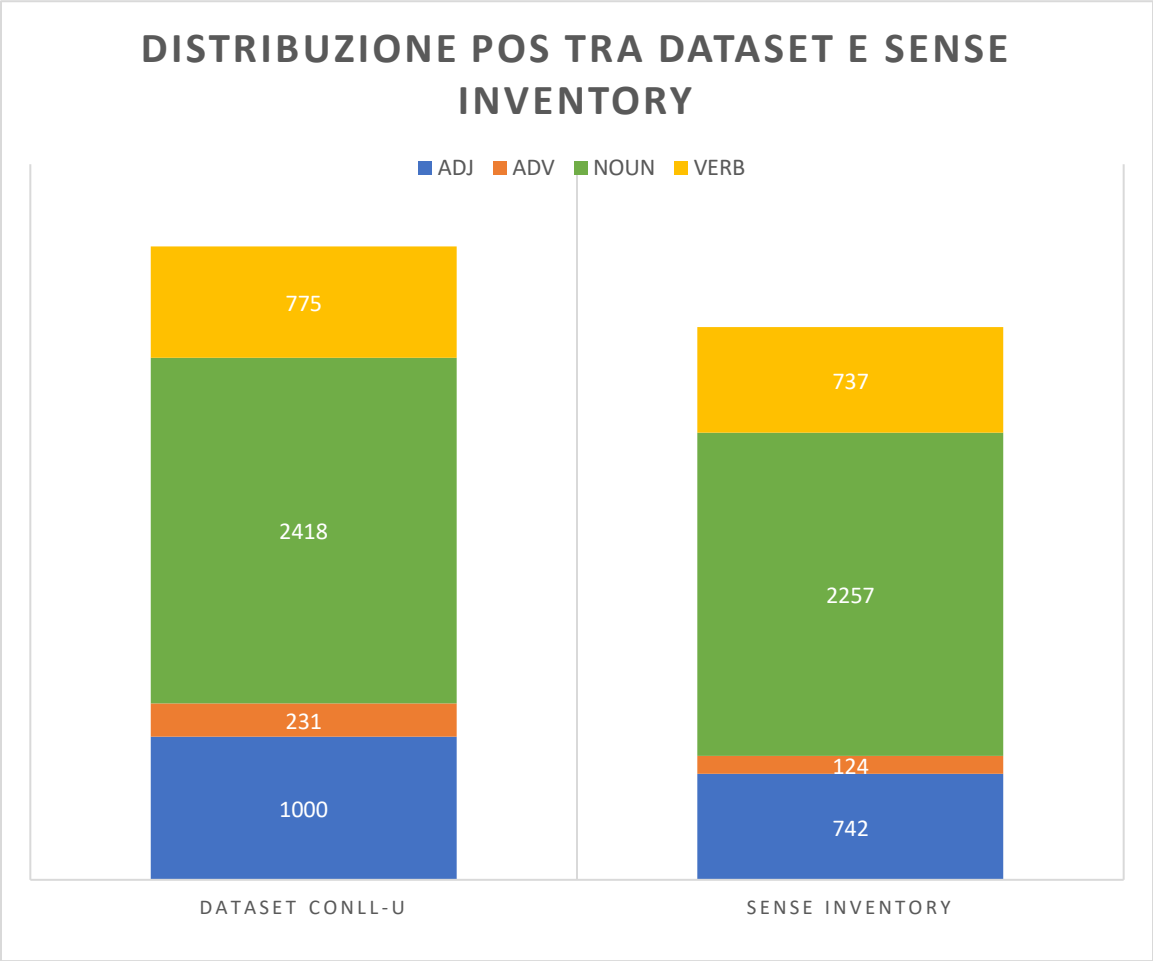

fig. 3: grafico a colonne che confronta la distribuzione delle PoS associate ai lemmi nel dataset e nel Sense Inventory

I lemmi del dataset presenti in IWN sono 3749, con una copertura del 84.7%; in PSC invece ne sono presenti 3060, quindi circa il 69.2%. Quest'ultimo dato deriva dal fatto che spesso in PSC i lemmi del dataset sono presenti ma non coperti con tutte le PoS ad essi associabili, riducendo di molto la copertura rispetto al set di frasi, dove un determinato lemma invece è spesso presente

correlato a diverse PoS a seconda del contesto sintattico in cui si trova. Considerando invece il numero di sensi corrispondenti a ogni coppia lemma – PoS trovata, ne contiamo un totale di 15,672, di cui 9638 trovati in IWN e 6034 in PSC. Il numero minore di sensi di PSC rispetto a IWN esplicita una caratteristica dei lessici già molto evidente anche dal task di annotazione manuale: mentre in PSC le definizioni sono più generiche e cercano di raggruppare più sensi sotto lo stesso lemma, in IWN sono parcellizzate in numerosi sensi molto più specifici, differenziando tutte le sfumature di significato. Per esempio, in IWN sono presenti più spesso sensi figurati o rari. Inoltre, dal momento che in PSC è stato reperito un numero inferiore di lemmi rispetto a IWN, ovviamente il numero di sensi sarà proporzionato alla percentuale di copertura.

Un'altra osservazione interessante, è che alcuni sensi sono privi di definizione, più precisamente 578 in IWN e 53 in PSC. Durante lo svolgimento del task manuale, si è ovviato a questo problema con un controllo sui *synset* o sul database di mapping, per poter scegliere come definizione o quella legata al *synset* più vicino o quella del senso mappato dall'altra risorsa, se presente.

L'intersezione tra i due lessici IWN e PSC conta 2949 lemmi con la stessa PoS, i cui sensi correlati sono di conseguenza potenzialmente mappabili se non già mappati. Questi sono diventati oggetto di test di mapping sulla base di similarità vettoriale della definizione, descritto nel capitolo 3. Un set di 911 lemmi, invece, è presente solo in una delle due risorse: di questi 800 appaiono solo in IWN; 111, solo in PSC: per questi ovviamente non è possibile espandere il mapping a meno di una futura integrazione dei database o di uno sviluppo che preveda un maggiore sfruttamento dei *synset*.

Di tutti i lemmi presenti nel dataset, quelli con almeno un senso mappato sono 2034, ovvero il 45.98%, mentre quelli non mappati in alcun senso sono 2389, ovvero il 54%; a questi ultimi, però, dobbiamo sottrarre i 911 non presenti sia in IWN che in PSC ma solo in una delle due risorse, per i quali è ovvio che non esistano mappature. In totale, i sensi mappati tra tutti i lemmi risultano essere 3461, ovvero solo il 22.1% dei sensi tra quelli in totale presenti nell'unione tra IWN e PSC.

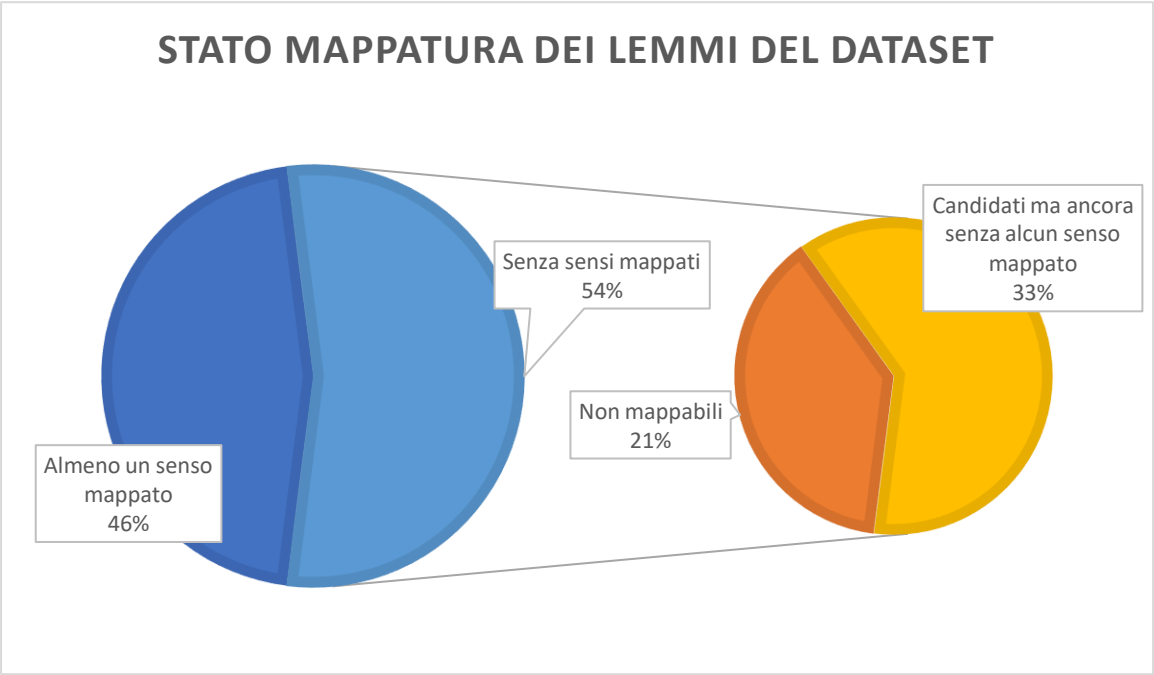

fig. 4: grafico a torta che illustra le casistiche di mappatura dei lemmi

5 Conclusioni

Il progetto di tesi presentato si pone nel paradigma della costruzione di risorse lessicali *ex novo* per riunire e armonizzare risorse pre-esistenti che siano di riferimento per la comunità di ricerca della lessicografica digitale.

Per raggiungere l'obiettivo, sono stati utilizzati strumenti e metodi che si avvalgono delle tecnologie computazionali per le materie umanistiche e in particolare per lo studio del linguaggio. Nel contesto del progetto ELEXIS, che mira a sviluppare una infrastruttura europea a vantaggio del settore della e-lexicography secondo gli obiettivi sopra citati, è stato sviluppato un Sense Inventory, ovvero una risorsa contenente un inventario di sensi correlati ad informazioni semantiche e destinato ad essere utilizzato per l'annotazione automatica o semi automatica di corpora paralleli multilingui. Nello specifico, il Sense Inventory frutto di questo lavoro è una risorsa lessicale in lingua italiana, che contribuirà alla codifica e all'annotazione della porzione italiana di un corpus multilingue all'interno del progetto ELEXIS. La metodologia proposta, tuttavia, non è specifica per le risorse dell'italiano, ma si pone come una procedura potenzialmente applicabile a tutte le lingue, universale, come richiesto dai princìpi e dagli obiettivi di ELEXIS. Infatti, migliorare lo sviluppo di lessici multilingue è di primaria importanza per la collaborazione interculturale; poiché essi sono la base di molte applicazioni multilingue e, oltre allo sviluppo del lessico, si ambisce anche all'arricchimento dei lessici monolingue di partenza, attraverso lo sfruttamento delle informazioni semantiche codificate in essi contenute (Bertagna, Monachini, Soria *et al.*, 2007).

Il Sense Inventory è stato creato attraverso il collegamento delle coppie lemma – PoS estrapolate dal dataset di partenza (in formato CoNLL-U), i cui sensi corrispondenti sono stati estratti dai due lessici fondamentali dell'italiano depositati nel repository ILC4CLARIN, ovvero ItalWordNet e PAROLE-SIMPLE-CLIPS. Il dataset corrisponde alla porzione italiana del corpus multilingue parallelo, codificata e rivista manualmente dal gruppo di ricerca italiano coinvolto nel progetto ELEXIS; il corpus contiene delle frasi selezionate dal web e l'elenco delle parole presenti in esse, corredate di altre informazioni quali ad esempio la forma di parola, il lemma corrispondente, la PoS, etc. Oltre ai sensi e le loro definizioni provenienti dai due lessici, sono stati integrati nel Sense Inventory anche i mapping tra sensi contenuti nel database IWNMAPDB, sviluppato come progetto dell'ILC-CNR con i vantaggi di una maggiore ricchezza delle definizioni e un maggior numero di informazioni semantiche.

Lavorando in questo modo, è stata ottenuta una risorsa che metta a disposizione un insieme completo delle informazioni necessarie al task di annotazione semantica di corpora, in maniera strutturata e aggiornata con i dati semantici indispensabili.

A questo compito principale sono stati aggiunti dei task di annotazione linguistica volti sia alla valutazione delle risorse sia al loro perfezionamento. Come parte sperimentale di questa tesi, è stato condotto un esperimento di estensione del mapping di sensi del database IWNMAPDB tramite l'utilizzo della libreria NLP spaCy. Attraverso gli algoritmi *word2vec* basati su *word embeddings* (vettori di parole) contenuti nelle funzioni di spaCy, è stato possibile realizzare un metodo per il calcolo di *similarity* (vicinanza semantica) tra le definizioni dei sensi correlati per reperire possibili candidati al mapping.

Infine, è stato creato un *test set* di frasi annotate manualmente a livello semantico, tramite il quale validare la correttezza delle informazioni estratte nei database e inserite nel Sense Inventory e valutarne l'effettiva copertura rispetto al dataset.

Da tutte le informazioni ricavate da osservazioni e test, con l'aiuto di un programma che estraesse dei dati per fini statistici, sono state tratte anche delle valutazioni finali sull'adeguatezza ed efficienza delle risorse, sia quelle preesistenti sia il Sense Inventory sviluppato nel contesto della tesi.

L'aggiornamento e la generazione di risorse lessicali sono un punto focale della ricerca lessicografica. È di fondamentale importanza che si producano lessici che possano rappresentare sempre più adeguatamente le sfaccettature di una lingua oggetto di studio, anche per affinare gli strumenti computazionali a disposizione, che necessitano di grandi quantità di dati per generare un'annotazione efficiente di corpora. Tali risorse, inoltre, divengono a loro volta oggetto di studio e di analisi o possono costituire un punto di riferimento per l'estrazione di informazioni linguistiche. Il mutuo collegamento e l'arricchimento di lessici contribuisce a indagare le necessità e i requisiti dell'integrazione e della capacità di scambi d'informazioni semi-automatica delle risorse lessicali, anche nei casi di *linking* in contesto multilingue e multiculturale (Soria, Monachini, Bertagna *et al.*, 2009). Nel presente caso di studio, invece, questi *task* sono stati svolti tra due lessici concepiti su modelli teorici diversi, ma per la stessa lingua.

Attraverso il lavoro svolto in questa tesi si è voluto dimostrare che la creazione di nuove risorse lessicali a partire da quelle preesistenti, oltre ad ottimizzare la loro struttura e copertura, serve a valutare lo stato dell'arte, contribuendo, al contempo, all'individuazione di nuovi metodi di

ampliamento e integrazione dei dati originari. Tali risorse, in formati che garantiscano l'interoperabilità, distribuite tramite infrastrutture lessicografiche pensate per l'open-source e la condivisione, si dimostrano di particolare interesse per la ricerca in un ambito come la e-lexicography, che è in rapida espansione nel mondo delle *digital humanities*.

Bibliografia

Articoli in rivista

Quochi, Valeria e Bartolini, Roberto e Monachini, Monica. 2017. *ItalWordNet goes open*. CSLI Publications. Linguistic Issues in Language Technology – LiLT, Vol. 10, Issue 4.

Ahmadi, Sina et al. 2020. *A Multilingual Evaluation Dataset for Monolingual Word Sense Alignment*. Marsiglia. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pag. 3232–3242. European Language Resources Association (ELRA), licensed under CC-BY-NC

Lenci, Alessandro et al. 2000. *SIMPLE: A General Framework for the Development of Multilingual Lexicons*. International Journal of Lexicography, Vol. 13, Issue 4, pag. 249–263

Roventini, Adriana e Ruimy, Nilda. 2008. *Mapping Events and Abstract Entities from PAROLE-SIMPLE-CLIPS to ItalWordNet*. Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, Marrakech, Morocco

Roventini, Adriana e Ruimy, Nilda e Marinelli, Rita e Ulivieri, Marisa e Mammini, Michele. 2007. *Mapping Concrete Entities from PAROLE-SIMPLE-CLIPS to ItalWordNet: Methodology and Results.* Proceedings of the ACL 2007 Demo and Poster Sessions, pag. 161–164, Prague, June 2007. ILC – CNR

Camacho-Collados, Jose e Pilehvar, Mohammad Taher. 2018. *From Word to Sense Embeddings: A Survey on Vector Representations of Meaning*. AI Access Foundation. Journal of Artificial Intelligence Research 63, pag. 743-788

Soria, Claudia e Monachini, Monica e Bertagna, Francesca e Calzolari, Nicoletta e Huang, Chu-Ren e Hsieh, Shu-Kai e Marchetti, Andrea e Tesconi, Maurizio. (2009). *Exploring interoperability of language resources: The case of cross-lingual semi-automatic enrichment of wordnets.* Language Resources and Evaluation. 43. 87-96. 10.1007/s10579-009-9082-3.

Bertagna, Francesca e Monachini, Monica e Soria, Claudia e Calzolari, Nicoletta e Huang, Chu-Ren e Hsieh, Shu-Kai e Marchetti, Andrea e Tesconi, Maurizio. (2007). *Fostering Intercultural Collaboration: A Web Service Architecture for Cross-Fertilization of Distributed Wordnets*. 146-158. 10.1007/978-3-540-74000-1_11.

Navigli, Roberto. 2009. *Word Sense Disambiguation: A Survey.* Università di Roma La Sapienza. *ACM Computing Surveys*, Vol. 41, No. 2, Articolo 10.

Volumi

Vossen P. (Ed.) (1998). *EuroWordNet: A multilingual database with lexical semantic networks.* Kluwer Academic.

Christiane Fellbaum (ed.) 1998. *Wordnet: An Electronic Lexical Database*. MIT Press.

Pustejovsky J. (1995). *The generative lexicon*. MIT Press.

Risorse

(1) Home repository ILC4CLARIN. https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/

(2) AA. VV., 2016, *PAROLE-SIMPLE-CLIPS*, ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa. http://hdl.handle.net/20.500.11752/ILC-88

(3) Roventini, Adriana; Marinelli, Rita and Bertagna, Francesca, 2016, *ItalWordNet v.2*, ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa, http://hdl.handle.net/20.500.11752/ILC-62 .

(4) *Python 3.8.3*. Data di release: 13 maggio 2020. Versione per Windows. https://www.python.org/downloads/release/python-383/

(5) Matias Grioni. 2018. *Pyconll. A minimal, all python, no dependency library to parse CoNLL files*. Licensed under the MIT License on GitHub. https://pyconll.github.io/

(6) Andy Dustman. 2012. *MySQLdb,* Revision 6c67620b. https://mysqlclient.readthedocs.io/

(7) Mike Kaplinskiy. *unicodedata*. Apache License 2.0. https://pypi.org/project/unicodedata2/

(8) Honnibal, Matthew and Montani, Ines and Van Landeghem, Sofie and Boyd, Adriane. 2020. *spaCy: Industrial-strength Natural Language Processing in Python*, Zenodo, 10.5281/zenodo.1212303, https://doi.org/10.5281/zenodo.1212303

(9) Repository GitHub del progetto. https://github.com/francescapoli98/thesis-project

(10) Sezione "Deliverables" del sito ufficiale di ELEXIS, con la documentazione riguardante gli obiettivi e gli sviluppi del progetto oltre che i report dei risultati ottenuti. https://elex.is/deliverables/

(11) Slides a cura del prof. Marcello Ferro, corso di Psicolinguistica Computazionale, a.a. 2019/2020

Appendice

Estratto dall'output del Sense Inventory: primi 25 lemmi

```
LEMMA POS   DEFINIZIONE CONCATENATA PSC-IWN    USEMID PSC  DEFINIZIONE PSC
      ESEMPIO PSC TIPO SEMANTICO PSC      SYNSETID IWN      SENSEID IWN
      DEFINIZIONE IWN

abbandonare VERB  cessare di fare qualcosa <> None    USem59592abbandonare
      cessare di fare qualcosa      Luca ha abbandonato l'attività
imprenditoriale   Cause_Aspectual   None  None  None
abbandonare VERB  lasciare <> None  USem73115abbandonare    lasciare
      abbandonare il paese, la moglie     Relational_Act    None  None  None
abbandonare VERB  far restare in un luogo <> None     USem79706abbandonare
      far restare in un luogo abbandonare la refurtiva in un campo
      Stative_Location  None  None  None
abbandonare VERB  None <> piegare, lasciare cadere una parte del corpo.None
      None  None  None  33547 46122 piegare, lasciare cadere una parte del
corpo.
abbandonare VERB  None <> lasciare definitivamente qualcuno, anche lasciare
senza aiuto, sostegno e sim.  None  None  None  None  33777 46530 lasciare
definitivamente qualcuno, anche lasciare senza aiuto, sostegno e sim.
abbandonare VERB  None <> desistere da un'impresa, cessare di fare qualcosa o
di prendersi cura di qualcosa.      None  None  None  None  33778 46533
```

desistere da un'impresa, cessare di fare qualcosa o di prendersi cura di qualcosa.
abbandonare VERB None <> lasciare andare qualcosa, abbandonando la presa.
None None None None 33779 46534 lasciare andare qualcosa, abbandonando la presa.
abbreviare VERB troncare con una abbreviazione; nella metrica classica, rendere breve una sillaba lunga <> rendere più breve.USem069533abbreviare
troncare con una abbreviazione; nella metrica classica, rendere breve una sillaba lunga abbreviare una parola Relational_Act 33384 44053 rendere più breve.
abbreviare VERB rendere più breve USem63040abbreviare rendere più breve
abbreviare i tempi di attesa Cause_Change 33384 44053 rendere più breve.
aberrazione NOUN anomalia, irregolarita' di organi e funzioni; aberrazione astigmatica, cromatica <> None USem74954aberrazione anomalia, irregolarita' di organi e funzioni; aberrazione astigmatica, cromatica questa legge è una vera aberrazione Constitutive None None None
aberrazione NOUN None <> stato patologico di alterazione della psiche.None
None None None 9541 14037 stato patologico di alterazione della psiche.
aberrazione NOUN None <> atto, azione aberrante. None None None None
9542 14039 atto, azione aberrante.
abile ADJ che ha i requisiti necessari per fare qlco <> None
USem62005abile che ha i requisiti necessari per fare qlco abile a svolgere un lavoro; - al lavoro PsychologicalProperty None None None
abile ADJ esperto, capace <> None USemD6670abile esperto, capace abile negli affari PsychologicalProperty None None None
abile ADJ None <> che è dichiarato idoneo ad un certo lavoro o servizio.
None None None None 41348 56223 che è dichiarato idoneo ad un certo lavoro o servizio.
abile ADJ None <> Che ha molta esperienza in un certo campo o della vita in generale None None None None 41349 56225 Che ha molta esperienza in un certo campo o della vita in generale
abile ADJ None <> detto di chi agisce con astuzia None None None None
41350 56230 detto di chi agisce con astuzia
abilità NOUN la qualità di chi è abile <> l'essere in grado di agire, di comportarsi in un certo modo, attitudine acquisita o innata a fare qualcosa.
USem73619abilita1 la qualità di chi è abile l'abilità di Leo di cucinare Quality 938 2095 l'essere in grado di agire, di comportarsi in un certo modo, attitudine acquisita o innata a fare qualcosa.
abitante NOUN chi abita in un luogo USemD659abitante chi abita in un luogo un abitante di Latina People 22292 27564 chi abita in un luogo.
abitare VERB occupare un luogo vivendoci; popolare <> vivere in un determinato luogo, avere lì la propria abitazione. USem76559abitare
occupare un luogo vivendoci; popolare molti animali feroci abitano la giungla Stative_Location 37863 50824 vivere in un determinato luogo, avere lì la propria abitazione.
abitare VERB risiedere, alloggiare <> vivere in un determinato luogo, avere lì la propria abitazione. USemD5057abitare risiedere, alloggiare
abitare al secondo piano Stative_Location 37863 50824 vivere in un determinato luogo, avere lì la propria abitazione.
abitativo ADJ relativo alle abitazioni o all'abitare <> None
USem75371abitativo relativo alle abitazioni o all'abitare locale ad uso abitativo ObjectRelated None None None
abito NOUN capo d'abbigliamento che si indossa sopra gli indumenti intimi <> qualsiasi capo di vestiario che si indossa sopra la biancheria. USem2963abito
capo d'abbigliamento che si indossa sopra gli indumenti intimi abito

da sera; abito da sposa Clothing 3189 5947 qualsiasi capo di vestiario che si indossa sopra la biancheria.
abito NOUN None <> tendenza acquisita che deriva dalla ripetizione costante di atti o comportamenti. None None None None 3190 5948 tendenza acquisita che deriva dalla ripetizione costante di atti o comportamenti.
abitudine NOUN consuetudine, tendenza acquisita <> tendenza acquisita che deriva dalla ripetizione costante di atti o comportamenti. USemD6398abitudine
consuetudine, tendenza acquisita fare qualcosa per abitudine; l'abitudine di Luca di dormire dopo pranzo Abstract_Entity 3190 5952
tendenza acquisita che deriva dalla ripetizione costante di atti o comportamenti.
abolizione NOUN l'abolire <> annullare qualcosa; l'annullare.
USem73621abolizione l'abolire l'abolizione della pena di morte da parte degli stati europei Cause_Change 2980 5619 annullare qualcosa; l'annullare.
aborto NOUN interruzione di gravidanza <> None USem74956aborto
interruzione di gravidanza essere contrario all'aborto
Non_Relational_Act None None None
aborto NOUN None <> interruzione di gravidanza None None None None
6762 11304 interruzione di gravidanza
aborto NOUN None <> persona o cosa brutta o imperfetta. None None
None None 6763 11305 persona o cosa brutta o imperfetta.
aborto NOUN None <> insuccesso, fallimento, cattiva riuscita di qualcosa. None None None None 6764 11306 insuccesso, fallimento, cattiva riuscita di qualcosa.
abuso NOUN uso esagerato e cattivo <> atto d'abuso; il prevaricare
USem74968abuso uso esagerato e cattivo abuso di fumo da parte di qualcuno Relational_Act 21379 28530 atto d'abuso; il prevaricare
abuso NOUN uso esagerato e cattivo <> uso eccessivo USem74968abuso uso esagerato e cattivo abuso di fumo da parte di qualcuno Relational_Act
21380 28532 uso eccessivo
accademia NOUN associazione di studiosi <> società per la difesa dei beni culturali e lo studio delle lettere USemD5255accademia associazione di studiosi accademia della Crusca Human_Group 1269 2664 società per la difesa dei beni culturali e lo studio delle lettere
accademia NOUN istituzione che ha come scopo l'incremento dello studio delle lettere, delle arti o delle scienze <> società per la difesa dei beni culturali e lo studio delle lettere USemD7028accademia istituzione che ha come scopo l'incremento dello studio delle lettere, delle arti o delle scienze
fare studi presso l'accademia Institution 1269 2664 società per la difesa dei beni culturali e lo studio delle lettere
accademia NOUN edificio in cui si riuniscono gli accademici o gruppi di studiosi <> istituto di insegnamento superiore USemD7029accademia
edificio in cui si riuniscono gli accademici o gruppi di studiosi
costruire un'accademia Building 1270 2665 istituto di insegnamento superiore
accademia NOUN None <> scuola di pensiero fondata da Platone None None
None None 1268 2662 scuola di pensiero fondata da Platone
accademia NOUN None <> trattenimento in cui si esibiscono gli allievi di un collegio None None None None 1271 2666 trattenimento in cui si esibiscono gli allievi di un collegio
accademia NOUN None <> dimostrazione di stile, specie riferita allo scherma e alla ginnastica None None None None 1272 2667 dimostrazione di stile, specie riferita allo scherma e alla ginnastica
accademia NOUN None <> studio o abbozzo fatto per esercizio copiando un modello None None None None 1273 2668 studio o abbozzo fatto per esercizio copiando un modello

accademia NOUN None <> esercitazione scolastica o retorica None None None None None 1274 2669 esercitazione scolastica o retorica
accademico ADJ relativo ad un'accademia artistica, letteraria o scientifica <> None USem75404accademico relativo ad un'accademia artistica, letteraria o scientifica comitato accademico ObjectRelated None None None
accademico ADJ dell'università <> None USem75410accademico dell'università senato accademico, calendario accademico ObjectRelated None None None
accademico ADJ che segue modelli artistici o letterari tradizionali, invalsi <> None USem75417accademico che segue modelli artistici o letterari tradizionali, invalsi opera accademica PsychologicalProperty None None None
accademico ADJ astratto e vuoto, ozioso, pomposo <> None USem75429accademico astratto e vuoto, ozioso, pomposo discorso accademico PsychologicalProperty None None None
accademico ADJ None <> Che è ozioso; che non raggiunge il suo scopo. None None None None 41458 56502 Che è ozioso; che non raggiunge il suo scopo.
accademico ADJ None <> concernente le accademie, quella di Platone e dei suoi successori in particolare None None None None 44166 59828 concernente le accademie, quella di Platone e dei suoi successori in particolare
accademico ADJ None <> Dell'Università None None None None 44167 59829 Dell'Università
accademico ADJ None <> Detto di studioso o di artista poco creativo per troppo rispetto dei modelli None None None None 44168 59830 Detto di studioso o di artista poco creativo per troppo rispetto dei modelli
accadere VERB capitare, succedere <> None USem3951accadere capitare, succedere accade sempre che piova in inverno Event None None None
accadere VERB None <> avere luogo, succedere.; avere luogo. None None None None 32413 42212 avere luogo, succedere.; avere luogo.
accampamento NOUN campo militare che provvede alloggi per soldati fatti da tende e baracche <> alloggiamento militare costitutito da tende o baracche. USem60244accampamento campo militare che provvede alloggi per soldati fatti da tende e baracche None Artifactual_area 16543 24012 alloggiamento militare costitutito da tende o baracche.
accampamento NOUN None <> l'accamparsi. None None None None 19117 30022 l'accamparsi.
accanto ADV None <> None None None None None 48710 64794 None
accecare VERB rendere privo dell'uso della ragione <> None USem62994accecare rendere privo dell'uso della ragione l'ira mi ha accecato Cause_Experience_Event None None None
accecare VERB abbagliare <> None USem63001accecare abbagliare quella luce mi ha accecato Cause_Change None None None
accecare VERB chiudere <> chiudere una porta o una finestra. USem63003accecare chiudere accecare una tubatura Cause_Change_of_State 35796 48043 chiudere una porta o una finestra.
accecare VERB rovinarsi la vista <> divenire cieco. USem63004accecare rovinarsi la vista accecarsi sui libri Change_of_State 35795 48041 divenire cieco.
accecare VERB privare della vista USem74633accecare privare della vista L'inquisitore accecò l'eretico Cause_Change_of_State 35794 48038 privare della vista.
accedere VERB Entrare <> None USemTH247accedere Entrare Max accede al forte Move None None None

accedere VERB Riuscire ad ottenere una carica <> None USemTH359accedere Riuscire ad ottenere una carica Accedere alla magistratura, alla carica di presidente Constitutive_Change None None None
accedere VERB Acconsentire <> None USemTH75accedere Acconsentire Max accede alle sue richieste, alla maggioranza (FIG) Relational_Act None None None
accedere VERB None <> entrare, avere accesso. None None None None 33796 46559 entrare, avere accesso.
accedere VERB None <> entrare a far parte di un complesso di organi o uffici. None None None None 33797 46560 entrare a far parte di un complesso di organi o uffici.
accedere VERB None <> avere accesso a una memoria o a un sistema informatico. None None None None 51213 68377 avere accesso a una memoria o a un sistema informatico.
accelerazione NOUN l'accelerare <> None USem60253accelerazione l'accelerare l'accelerazione della macchina Change None None None
accelerazione NOUN None <> variazione della velocità nell'unità di tempo None None None None 21694 26902 variazione della velocità nell'unità di tempo
accelerazione NOUN None <> l'accelerare. None None None None 21005 28099 l'accelerare.
accennare VERB far pensare, far prevedere <> None USem75163accennare far pensare, far prevedere La pioggia non accenna a smettere Event None None None
accennare VERB dar l'idea di voler fare qualcosa <> fare l'atto di. USem75161accennare dar l'idea di voler fare qualcosa accennare un sorriso, uno schiaffo Act 34605 44305 fare l'atto di.
accennare VERB fare cenno, indicare <> esprimersi mediante cenni. USem75162accennare fare cenno, indicare accennare con lo sguardo, con la mano Relational_Act 34604 44304 esprimersi mediante cenni.
accennare VERB fare cenno, indicare <> fare l'atto di. USem75162accennare fare cenno, indicare accennare con lo sguardo, con la mano Relational_Act 34605 44305 fare l'atto di.
accennare VERB fare cenno, indicare <> indicare q.c.a qc. USem75162accennare fare cenno, indicare accennare con lo sguardo, con la mano Relational_Act 34608 44309 indicare q.c.a qc.
accennare VERB alludere, parlare in modo approssimativo <> parlare brevemente e superficialmente (fig.). USem75164accennare alludere, parlare in modo approssimativo accennare a un problema Reporting_Event 34606 44307 parlare brevemente e superficialmente (fig.).
accennare VERB alludere, parlare di sfuggita <> parlare brevemente e superficialmente (fig.). USem75165accennare alludere, parlare di sfuggita Ha accennato il problema; - che ci saranno dei problemi; - di aver saputo dei problemi; alla popolazione Reporting_Event 34606 44307 parlare brevemente e superficialmente (fig.).
accennare VERB None <> dare indizio di, cominciare. None None None None 34607 44308 dare indizio di, cominciare.
accennare VERB None <> dire, in modo velato. None None None None 33943 46839 dire, in modo velato.
accento NOUN segno grafico <> None USem1703accento segno grafico None Sign None None None
accento NOUN messa in rilievo di una sillaba <> segno diacritico usato per evidenziare la sillaba tonica USem68928accento messa in rilievo di una sillaba accento sulla penultima sillaba Metalanguage 30872 39937 segno diacritico usato per evidenziare la sillaba tonica

accento NOUN modo di pronunciare <> intensità che la voce assume su una sillaba di una parola USem68930accento modo di pronunciare avere l'accento francese Quality 30873 39938 intensità che la voce assume su una sillaba di una parola
accento NOUN None <> modo di modulare una voce o un suono, spesso tipico di alcune regioni None None None None 7736 12658 modo di modulare una voce o un suono, spesso tipico di alcune regioni
accentuare VERB accrescersi <> None USem6810accentuare accrescersi la crisi monetaria si e' accentuata Change_of_Value None None None
accentuare VERB aggravare <> far aumentare. USem6752accentuare aggravare questo fatto ha accentuato la crisi monetaria Cause_Change_of_Value 40508 55345 far aumentare.
accentuare VERB None <> segnare con l'accento. None None None None 37976 50972 segnare con l'accento.
accentuare VERB None <> pronunciare con enfasi. None None None None 40505 55342 pronunciare con enfasi.
accentuare VERB None <> dare maggior rilievo. None None None None 40506 55343 dare maggior rilievo.
accentuare VERB None <> porre in evidenza. None None None None 40507 55344 porre in evidenza.
accertamento NOUN l'accertare <> indagine investigativa su una certa materia; indagine investigativa; atto del verificare.; atto del verificare; l'atto dell'esaminare; l'esaminare USem069743accertamento l'accertare l'accertamento dei fatti da parte degli inquirenti Acquire_Knowledge 2156 4181 indagine investigativa su una certa materia; indagine investigativa; atto del verificare.; atto del verificare; l'atto dell'esaminare; l'esaminare
accertamento NOUN l'accertare <> verifica di prove e situazioni all'interno di una investigazione legale USem069743accertamento l'accertare l'accertamento dei fatti da parte degli inquirenti Acquire_Knowledge 15463 22713 verifica di prove e situazioni all'interno di una investigazione legale
accertamento NOUN l'accertare <> attività diretta ad eliminare una situazione giuridica incerta. USem069743accertamento l'accertare l'accertamento dei fatti da parte degli inquirenti Acquire_Knowledge 15464 22714 attività diretta ad eliminare una situazione giuridica incerta.
accessibile ADJ comprensibile <> None USem75446accessibile comprensibile testo accessibile ai più PsychologicalProperty None None None
accessibile ADJ a cui si può accedere <> None USem75455accessibile a cui si può accedere locale accessibile ai disabili Modal None None None
accessibile ADJ modico <> None USem75459accessibile modico prezzo accessibile a tutti SocialProperty None None None
accessibile ADJ disponibile <> None USem75466accessibile disponibile persona accessibile PsychologicalProperty None None None
accessibile ADJ None <> non troppo caro nel prezzo. None None None None 42645 57498 non troppo caro nel prezzo.
accessibile ADJ None <> che si può raggiungere facilmente. None None None None 43078 58239 che si può raggiungere facilmente.