UNIVERSITÀ DEGLI STUDI DI SALERNO

Dipartimento di
Scienze Politiche e della Comunicazione

Corso di Dottorato in
SCIENZE E CULTURE DELL'UMANO.
SOCIETA', EDUCAZIONE, POLITICA E COMUNICAZIONE
Curriculum: F – Politica, Diritto, Mercato
XXXVII Ciclo

TESI DI DOTTORATO

RISORSE LINGUISTICHE DIGITALI PER LA TRANSIZIONE VERDE

Relatrice:
Prof.ssa Daniela VELLUTINO

Candidato
Dott. Nicola CIRILLO

Matr: 8803700024

Coordinatore del dottorato
Chiar.mo Prof. Filippo FIMIANI

Anno Accademico 2023/2024

UNIONE EUROPEA
Fondo Sociale Europeo
Ministero dell'Università e della Ricerca
PON RICERCA E INNOVAZIONE
REACT EU

Desidero rivolgere un sentito ringraziamento alla mia tutor, la Prof.ssa Daniela Vellutino, che mi ha introdotto con passione ai temi di ricerca che da anni rappresentano il fulcro del suo impegno scientifico e civile. Il framework teorico e la metodologia di analisi da lei elaborati hanno costituito una solida base su cui ho potuto sviluppare il mio lavoro di ricerca.

Un ringraziamento speciale va anche alla Prof.ssa Simonetta Vietri, per il suo costante supporto scientifico, per avermi introdotto allo studio della linguistica e per avermi incoraggiato a proseguire in questo ambito. La sua esperienza e competenza sono state fondamentali.

Abstract (*Italiano*)

Nel lessico dell'italiano istituzionale, la terminologia specialistica svolge il ruolo fondamentale di esprimere concetti complessi relativi alle diverse sfere d'azione delle istituzioni. Di conseguenza, può determinare, anche da sola, il successo (o l'insuccesso) della comunicazione tra le istituzioni e il loro pubblico di riferimento. Perciò, l'utilizzo corretto della terminologia favorisce il buon funzionamento delle istituzioni. Esemplare è il tema della gestione dei rifiuti urbani, dove la corretta comunicazione è uno dei fattori determinanti per raggiungere gli obiettivi ambientali e di raccolta differenziata fissati dalla normativa europea.

In questo senso, la terminologia istituzionale relativa alla gestione dei rifiuti urbani presenta alcune criticità che, di fatto, ostacolano la corretta comunicazione tra istituzioni e pubblico. Il primo problema è sicuramente l'alto grado di variabilità dei termini utilizzati per indicare le frazioni oggetto di raccolta differenziata. Basti pensare che, per il solo rifiuto urbano residuo, abbiamo individuato ben 18 denominazioni (*indifferenziato*, *secco residuo*, *secco non riciclabile*, ecc.) e per i rifiuti biodegradabili almeno 15 (*umido*, *organico*, *residuo putrescibile*, ecc.). Un'altra criticità è rappresentata dall'uso di termini ad alto e altissimo specialismo in contesti comunicativi che coinvolgono un pubblico generalista. Ad esempio, segnaliamo l'utilizzo dell'acronimo *CCR* in alcuni opuscoli, per indicare il *Centro Comunale di Raccolta* o, ancora, il termine *scarrabile* per indicare il cassone dove si depositano i rifiuti all'interno del centro di raccolta. Infine, ci sono alcuni casi in cui le istituzioni (o chi comunica per loro) coniano autonomamente nuovi termini che sono poco trasparenti o addirittura fuorvianti. Ad esempio, il termine *sacchetto a svuotamento* sembra indicare uno specifico tipo di sacchetto ma, in realtà, indica semplicemente che non bisogna gettare il sacchetto nel bidone ma bensì svuotarne il contenuto (come avviene, ad esempio, nei cestini per le pile esauste).

Riteniamo che, per affrontare questi problemi, un approccio prescrittivo volto alla sola standardizzazione terminologica non sia sufficientemente efficace. Infatti, tale approccio non prende in considerazione le specifiche esigenze delle diverse situazioni comunicative. È quindi necessario integrarlo con un approccio di tipo descrittivo, orientato a produrre risorse terminologiche di riferimento che possano, da un lato, guidare una standardizzazione più consapevole e dall'altro aiutare i comunicatori pubblici a destreggiarsi meglio nel complesso panorama della terminologia istituzionale e scegliere la variante terminologica più appropriata ad esprimere un dato concetto in una determinata situazione comunicativa.

A questo scopo, abbiamo analizzato la terminologia istituzionale della gestione dei rifiuti urbani e realizzato quattro risorse linguistico-terminologiche: Il corpus ItaIst-DdAC_GRU, una raccolta di testi istituzionali speciali e mediali sul tema della gestione dei rifiuti urbani; il corpus ItaIst-TermRifiuti, composto da un campione di ItaIst-DdAC_GRU in cui la terminologia è stata annotata manualmente da quattro annotatori; la tecnica di estrazione automatica di termini *Domain Con-*

cept Relatedness, focalizzata sull'isolare la terminologia di uno specifico dominio e, infine, la banca dati terminologica ItaIst-WasteLexicon che contiene 922 termini relativi alla gestione dei rifiuti organizzati secondo una struttura concettuale gerarchica.

Abstract (*English*)

In the lexicon of institutional Italian, specialized terminology plays a fundamental role in expressing complex concepts related to institutions' various fields of action. Consequently, it can determine, even on its own, the success or failure of communication between institutions and their target audience. Therefore, the correct use of terminology promotes the effective functioning of institutions.

A prime example is the topic of municipal waste management, where accurate communication is crucial for achieving environmental and separate collection goals set by European legislation.

In this regard, institutional terminology related to municipal waste management presents some critical issues that hinder effective communication between institutions and the public. The primary issue is undoubtedly the high variability in terms used to indicate waste categories. Just consider that, for residual urban waste alone, we have identified 18 different terms (e.g., "non-recyclable dry waste", "residual dry waste", etc.), and for biodegradable waste, at least 15 terms (e.g., "organic waste", "putrescible residue", etc.).

Another critical aspect is the use of highly specialized terms in communication contexts that involve a general audience. For instance, the acronym *CCR* is used in some brochures to refer to the *Centro Comunale di Raccolta* (Municipal Collection Center), or the term *scarrabile* that refers to the container where waste is deposited within the collection centre.

Lastly, there are cases where institutions (or those communicating on their behalf) coin new terms that are unclear or even misleading. For example, the term *sacchetto a svuotamento* (emptying bag) seems to indicate a specific type of bag but actually means that one should empty its contents rather than discard the bag itself (as seen, for example, in bins for exhausted batteries).

To address these issues, we believe that a prescriptive approach focused solely on terminological standardization is not sufficiently effective. Such an approach overlooks the specific needs of different communicative situations. Therefore, it is necessary to integrate it with a descriptive approach aimed at producing reference terminological resources. These resources can guide more informed standardization and help public communicators navigate the complex landscape of institutional terminology, selecting the most appropriate variant to express a given concept in a specific communication context.

To this end, we have analyzed institutional terminology related to urban waste management and developed four linguistic-terminological resources: the ItaIst-DdAC_GRU corpus, a collection of specialized institutional and media texts on urban waste management; the ItaIst-TermRifiuti corpus, comprising a sample from ItaIst-DdAC_GRU with manually annotated terminology by four annotators; the Domain Concept Relatedness, an automatic term extraction technique, focused on isolating domain-specific terminology; and finally, the ItaIst-WasteLexicon terminological database containing 922 terms related to waste management organized in a hierarchical conceptual structure.

Indice

1 Introduzione **1**

1.1 L'italiano istituzionale 1
1.2 Terminologia istituzionale della gestione dei rifiuti 2
1.3 Risorse linguistiche come dati aperti 5
1.4 Organizzazione del lavoro 6

2 Quadro teorico **9**

2.1 Principali approcci alla terminologia 9
2.1.1 Teoria Classica 10
2.1.2 Approccio sociolinguistico (socioterminologia) 10
2.1.3 Approccio comunicativo 11
2.1.4 Approccio sociocognitivo 12
2.1.5 Approccio culturale 12
2.1.6 Approccio testuale 13
2.2 Sfondo teorico 14
2.2.1 Concetto di termine 14
2.2.2 Termini come unità lessicali 15
2.2.3 Le componenti del termine 17
2.2.4 Proprietà morfosintattiche del termine 18
2.2.5 Proprietà concettuali del termine 19
2.2.6 Proprietà semantico-comunicative del termine 22

3 Terminologia computazionale **23**

3.1 Introduzione 23
3.2 Corpora specialistici e di dominio 24
3.2.1 Corpora in terminologia 25
3.2.2 Corpora specialistici primari 26
3.2.3 Corpora specialistici secondari 27
3.2.4 Formati di rappresentazione 28
3.3 Estrazione automatica di termini 30
3.3.1 Approcci tradizionali 30
3.3.2 Approcci basati su *deep learning* e *prompting* 31
3.3.3 Tecniche per isolare uno strato terminologico 32
3.4 Gestione e pubblicazione delle risorse terminologiche 33

3.4.1 Formati di rappresentazione 33
3.4.2 Principali banche dati terminologiche istituzionali 37

4 Corpora **39**
4.1 Introduzione 39
4.2 ItaIst-DdAC_GRU 40
4.2.1 Raccolta e metadatazione dei testi 40
4.2.2 Estrazione del testo e preprocessing 42
4.2.3 Composizione del corpus 42
4.3 ItaIst-TermRifiuti 43
4.3.1 Accordo tra annotatori 45
4.4 Formati di rappresentazione e modalità di interrogazione 47

5 Domain Concept Relatedness **51**
5.1 Isolare la terminologia di dominio 51
5.2 Key Concept Relatedness 52
5.3 Domain Concept Relatedness 52
5.4 Sperimentazione 53
5.4.1 Corpora 53
5.4.2 Dataset 54
5.4.3 Metodologia di valutazione 54
5.4.4 Risultati 55

6 Base dati terminologica ItaIst-Waste Lexicon **57**
6.1 Introduzione 57
6.2 Metodologia di costruzione 58
6.3 Dimensione morfosintattica 59
6.3.1 Struttura morfologica dei termini monorematici 59
6.3.2 Struttura sintattica dei termini polirematici 60
6.4 Dimensione concettuale 61
6.5 Dimensione semantico-comunicativa 62
6.5.1 Rilevanza terminologica 63
6.5.2 Risultati 64
6.6 Formati di rappresentazione e modalità di interrogazione 65

7 Conclusioni **67**

Bibliografia **69**

Appendice A - Esempio di scheda terminologica **79**

Appendice B - Gradi di specialismo in ItaIst-DdAC_GRU **81**

Capitolo 1

Introduzione

Contenuti

1.1 L'italiano istituzionale **1**
1.2 Terminologia istituzionale della gestione dei rifiuti **2**
1.3 Risorse linguistiche come dati aperti **5**
1.4 Organizzazione del lavoro **6**

1.1 L'italiano istituzionale

La varietà di lingua utilizzata in Italia dalle istituzioni, denominata da Berruto (1987) "italiano burocratico", è da tempo oggetto di aspre critiche a causa della sua scarsa leggibilità, che interferisce con il diritto all'informazione (Cortelazzo, 2021; Lubello, 2014b; Vellutino, 2018). Nel tentativo di semplificare questa varietà di lingua, sono stati redatti diversi manuali e linee guida che contengono una serie di regole di riscrittura (Cortelazzo, 2021; Fiorentino & Ganfi, 2024; Fioritto, 1997; Lubello, 2014b; Piemontese, 1999; Vellutino, 2018). Nonostante i numerosi tentativi, ad oggi, il problema è ancora lontano dall'essere risolto Lubello, 2018.

Tuttavia, le varietà dell'italiano amministrativo e burocratico, descritte da Berruto (1987) e Lubello (2014a), hanno risentito dell'influenza dei processi di semplificazione, dell'*euro-english* e dei diversi mezzi di comunicazione, quali la stampa, Internet e i social media, e stanno progressivamente ridefinendosi in una nuova varietà linguistica, denominata da Vellutino (2018, 2022) italiano istituzionale, una varietà che tende verso la chiarezza e la semplicità, a differenza dell'italiano burocratico.

In questo lavoro si assume il paradigma di ricerca e la metodologia per lo studio degli usi dell'italiano istituzionale e si adotta la seguente nozione di italiano istituzionale (Vellutino, 2018, p. 45).

> L'italiano istituzionale è una varietà della lingua nazionale che si sta attestando attraverso gli usi linguistici, prevalentemente scritti, delle comunicazioni ufficiali delle istituzioni, delle amministrazioni pubbliche e, più in generale, delle organizzazioni del settore pubblico e privato.
> È una varietà linguistica funzionale, dipendente dai molteplici domini di conoscenza e

> dai vari settori di attività delle organizzazioni ed è usata da diversi gruppi di parlanti/scriventi per governare, legiferare, dirigere, amministrare, svolgere transazioni e comunicare al loro interno, con le altre istituzioni e con il pubblico distinto o indifferenziato che ha differenti livelli socioculturali e abilità linguistiche e comunicative.

Dall'analisi dei diversi usi dell'italiano istituzionale nelle amministrazioni pubbliche è emerso che questa varietà linguistica è impiegata in diversi contesti socio-pragmatici, che vanno dai linguaggi speciali del diritto e dell'amministrazione a quelli mediali per la comunicazione pubblica (Vellutino, 2018; Vellutino et al., 2012). Ai fini di questa trattazione, raggrupperemo gli usi dell'italiano istituzionale in due macro-categorie: i linguaggi istituzionali speciali e i linguaggi istituzionali mediali (Vellutino, 2018; Vellutino et al., 2012).

I linguaggi istituzionali speciali sono utilizzati dalle amministrazioni pubbliche per la comunicazione tra esperti. In particolare sono attestati nei documenti ufficiali quali atti legislativi e atti amministrativi. Invece, i linguaggi istituzionali mediali sono utilizzati nella comunicazione tra esperti e non esperti. Di conseguenza, si trovano attestati in brochure informative, comunicati stampa, news sui siti web istituzionali e post sui social media.

1.2 Terminologia istituzionale della gestione dei rifiuti

Nell'italiano istituzionale, la terminologia svolge un ruolo chiave: designare quei concetti che sono propri dei diversi settori di cui le istituzioni si occupano. Perciò, per migliorare la qualità dell' italiano istituzionale bisognerebbe innanzitutto analizzarne a fondo la terminologia specialistica e apprendere come impiegarla al meglio.

Per illustrare l'importanza della terminologia nella comunicazione pubblica istituzionale, esaminiamo alcuni estratti di un avviso pubblico relativo alla raccolta differenziata di pannolini e pannoloni, incluso nei corpora ItaIst-DdAC_GRU e ItaIst-Termrifiuti (cfr. Capitolo 4).

> Per questo motivo, la SARIM ha ricevuto disposizioni per l'installazione immediata sul territorio comunale di **[1] stazioni di conferimento automatizzate** per il conferimento di pannolini e pannoloni [...]
>
> Gli utenti abilitati possono conferire questa tipologia di rifiuti nella **[2] stazione ecologica** individuata per ognuno sulla base della dimora.[...]
>
> Nei **[3] contenitori** per pannolini e pannoloni vanno conferiti esclusivamente: - pannolini per bambini; - pannoloni per adulti e traverse. [...]

Il primo problema che emerge da questo avviso pubblico è la mancanza di coerenza terminologica. Infatti, i termini [1], [2] e [3], sebbene diversi, si riferiscono allo stesso concetto, quello di *cassonetto intelligente*,[1] che definiremo come "cassonetto dotato di sensori che permettono il riconoscimento dell'utente." L'uso di tre termini diversi per designare lo stesso concetto potrebbe confondere il lettore in quanto lascia aperta la possibilità che ogni termine abbia un diverso referente.

Inoltre, tutti e tre i termini sono ambigui e poco trasparenti. Nello specifico, i termini [1] e [2] hanno come testa il sostantivo *stazione*, ma, nel lessico comune, la parola *stazione* non indica un contenitore per gettare i rifiuti. Perciò, entrambi i termini necessitano di un modificatore che

[1] Il termine *cassonetto intelligente* è un calco di traduzione dall'inglese *smart bin*.

ne specifichi la funzione. Il risultato è una complessità e una lunghezza eccessive che riducono la leggibilità complessiva del testo. Anche il termine [3], sebbene sia più trasparente, rimane comunque troppo generico. Infatti, il termine *contenitore* è un iperonimo e, nel lessico della gestione dei rifiuti, viene usato spesso per indicare un bidone o un mastello e solo raramente è usato per riferirsi a un cassonetto.

Proviamo invece ad utilizzare, nello stesso testo, un termine più breve e trasparente: il termine *cassonetto intelligente*.

> Per questo motivo, la SARIM ha ricevuto disposizioni per l'installazione immediata sul territorio comunale di **cassonetti intelligenti** per il conferimento di pannolini e pannoloni [...]
>
> Gli utenti abilitati possono conferire questa tipologia di rifiuti nel **cassonetto intelligente** individuato per ognuno sulla base della dimora.[...]
>
> Nei **cassonetti intelligenti** per pannolini e pannoloni vanno conferiti esclusivamente: - pannolini per bambini; - pannoloni per adulti e traverse.[...]

Noteremo che questo termine, rispetto ai precedenti, rende più scorrevole il testo e più chiaro l'oggetto denotato che è, di fatto, un cassonetto. Oltretutto, la funzione di un cassonetto è insita nell'oggetto stesso e non necessita quindi di ulteriori specificazioni.

Questo lavoro affronta il tema della terminologia istituzionale della gestione dei rifiuti urbani, un settore in cui la comunicazione è determinante per raggiungere gli obiettivi ambientali e di raccolta differenziata fissati dalla normativa europea. Come abbiamo visto nell'esempio precedente, tale terminologia non è scevra da problematiche che ostacolano la comunicazione tra le istituzioni e i loro pubblici.

Una di queste problematiche è sicuramente la proliferazione dei termini che denotano le frazioni oggetto di raccolta differenziata. Il solo rifiuto urbano residuo presenta, nella banca dati terminologica che abbiamo sviluppato, ben 18 denominazioni diverse (*indifferenziato*, *secco residuo*, *secco non riciclabile*, ecc.) e i rifiuti biodegradabili ne presentano almeno 15 (*umido*, *organico*, *residuo putrescibile*, ecc.).

Un'altra criticità è rappresentata dall'uso di termini di alto e altissimo specialismo impiegati in contesti comunicativi che coinvolgono un pubblico generalista. Ad esempio, segnaliamo l'utilizzo dell'acronimo *CCR* in alcuni opuscoli, per indicare il *Centro Comunale di Raccolta* o, ancora, l'aggettivo sostantivato *scarrabile* per indicare il *cassone scarrabile* dove si depositano i rifiuti all'interno del centro di raccolta.

Infine, in alcuni casi, le istituzioni, o chi comunica per loro, coniano autonomamente nuovi termini che sono poco trasparenti o addirittura fuorvianti, come nel caso del termine *sacchetto a svuotamento* che sembrerebbe indicare uno specifico tipo di sacchetto ma, in realtà, indica semplicemente che non bisogna gettare il sacchetto nel bidone ma svuotarne il contenuto all'interno (come avviene, ad esempio, nei cestini per le pile esauste).

Per affrontare le problematiche appena illustrate, una strada possibile è quella della standardizzazione terminologica. Organismi come l' *International Standard Organizzation* (ISO) e, per l'Italia, l'Ente Italiano di Normazione (UNI), potrebbero adottare norme tecniche che regolino l'uso dei termini nel settore della gestione dei rifiuti urbani. Noi però riteniamo che quest'approccio, sebbene efficace, non sia da solo sufficiente. Il rischio è che tale approccio perda di vista le specifiche esigenze delle diverse situazioni comunicative.

L'approccio prescrittivo della standardizzazione, quindi, deve essere necessariamente affiancato ad un approccio descrittivo che guidi la standardizzazione e, al contempo, fornisca strumenti di consultazione utili a formare i comunicatori pubblici. Ovviamente, un'accurata descrizione della terminologia istituzionale della gestione dei rifiuti urbani deve fondarsi su una base empirica solida.

Quindi, per affrontare le problematiche sopra descritte, il presente lavoro si pone l'obiettivo di realizzare quattro risorse linguistico-terminologiche relative al dominio della gestione dei rifiuti urbani: due corpora, una banca dati terminologica e una tecnica di Estrazione Automatica di Termini (EAT). Queste risorse hanno almeno tre scopi. Il primo scopo, il principale, è sicuramente quello di fornire una base empirica per lo studio approfondito della terminologia istituzionale della gestione dei rifiuti urbani. Il secondo scopo è dotare i comunicatori pubblici di strumenti di consultazione che li aiutino a scegliere correttamente i termini, in base al pubblico di riferimento e alla situazione comunicativa. Il terzo scopo è fornire materiale di qualità per l'addestramento dei sistemi di Intelligenza Artificiale, che sono oggi sempre più diffusi, in modo da renderli capaci di trattare efficacemente i linguaggi istituzionali e la loro complessa terminologia.

Il presente lavoro si colloca quindi nell'ambito disciplinare degli studi linguistici e terminologici. Nello specifico, la prospettiva adottata è quella della terminologia computazionale, una disciplina focalizzata sullo sviluppo di strumenti informatici che servono a studiare la terminologia specialistica e a trattarla automaticamente. Tra questi strumenti troviamo, tra gli altri, corpora specialistici, banche dati terminologiche e algoritmi di EAT.

Il presente lavoro di tesi è esito di un percorso di ricerca condotto nell'ambito di un dottorato industriale del programma PON REACT-EU - Dottorati di ricerca innovazione e green. L'indagine si colloca all'interno di un consolidato filone di studi dedicati all'analisi degli usi dell'italiano istituzionale, sviluppatosi attraverso molteplici approcci linguistici e avvalorato da evidenze empiriche (Vellutino, 2014, 2018; Vellutino et al., 2012). In tale cornice teorica e metodologica, la ricerca qui presentata adotta il framework per l'analisi della costruzione e della distribuzione della terminologia specialistica all'interno dei testi istituzionali, applicandolo al dominio specifico della Gestione del servizio di raccolta differenziata dei rifiuti solidi urbani (Vellutino, 2015; Vellutino & Formisano, 2016; Vellutino et al., 2016).

Questa tesi si inserisce dunque in una linea di continuità con le ricerche precedenti e ne costituisce un'ulteriore articolazione, contribuendo a verificare l'ipotesi dell'esistenza di processi di variazione terminologica all'interno dei testi riconducibili ai diversi linguaggi istituzionali, in particolare quelli della comunicazione specialistica giuridico-amministrativa e della comunicazione pubblica rivolta a cittadini con differenti livelli di competenza linguistico-comunicativa (Cirillo & Vellutino, 2024; Vellutino, 2018; Vellutino & Cirillo, 2024).

L'obiettivo principale dello studio è la costruzione di risorse linguistiche e terminologiche attraverso una metodologia rigorosa fondata sulla letteratura specialistica. Il primo risultato significativo è rappresentato dall'elaborazione di un campionamento stratificato non proporzionale di testi istituzionali, selezionati sulla base della classificazione proposta in precedenti lavori (Vellutino, 2015, 2018; Vellutino et al., 2012). Tale attività ha condotto alla costruzione del corpus annotato ItaIst-TermRifiuti, derivato dal corpus ItaIst-DdAC_GRU. Dal corpus ItaIst-TermRifiuti sono stati estratti i termini specialistici, successivamente organizzati all'interno della banca dati terminologica ItaIst-WasteLexicon.

In sintesi, le risorse linguistico-terminologiche[2] che abbiamo realizzato e che saranno presentate

[2]Si precisa che le risorse linguistico-terminologiche sviluppate nel presente lavoro di ricerca sono state ideate

nel corso della trattazione sono le seguenti:

- Il corpus **ItaIst-DdAC_GRU** (Italiano Istituzionale per il Diritto di Accesso Civico del dominio terminologico Gestione Rifiuti Urbani): un corpus di monitoraggio che raccoglie testi istituzionali (atti amministrativi e testi informativi) riguardanti il tema della gestione dei rifiuti urbani.
- Il corpus **ItaIst-TermRifiuti**: un campione di ItaIst-DdAC_GRU in cui quattro annotatori hanno identificato manualmente la terminologia dei domini del diritto e dell'amministrazione e della gestione dei rifiuti urbani.
- La tecnica di EAT denominata **Domain Concept Relatedness**: una tecnica che, a differenza delle tecniche di EAT tradizionali, non estrae indiscriminatamente tutti i termini ma si focalizza sull'isolare la terminologia di un determinato dominio di conoscenza.
- La banca dati terminologica **ItaIst-WasteLexicon**: una raccolta di termini istituzionali della gestione dei rifiuti urbani orientata alla variazione terminologica, in cui il grado di specialismo dei termini è stato determinato empiricamente tramite l'analisi della loro distribuzione nei diversi tipi di testo.

ItaIst-DdAC_GRU, ItaIst-TermRifiuti e ItaIst-Waste lexicon sono distribuiti su Zenodo[3] con licenza CC-BY 4.0 (Cirillo et al., 2025). Invece, l'implementazione della DCR è distribuita su GitHub[4] con licenza MIT

Inoltre, lo sviluppo di queste risorse ha permesso un'analisi preliminare della terminologia istituzionale della gestione dei rifiuti urbani, le cui caratteristiche saranno quindi analizzate e discusse nel corso della trattazione.

L'attenzione al dominio terminologico, di fondamentale importanza per lo studio dell'italiano istituzionale, si traduce all'interno del lavoro nella proposta del modello a Strati Terminologici, nello sviluppo di una tecnica di EAT, la Domain Concept Relatedness, orientata ad isolare i termini di un dato dominio e nell'annotazione del corpus ItaIst-TermRifiuti con etichette di dominio. Invece, il focus sulla variazione terminologica e il grado di specialismo si è sostanziato nella scelta di metadatare il corpus ItaIst-DdAC GRU in base alla tipologia testuale, nel bilanciare il corpus ItaIst-TermRifiuti per tipo di testo e nel successivo sviluppo di una metrica in grado di catturare il grado di specialismo delle varianti terminologiche.

Infine, un'attenzione particolare è posta al tema dell'accesso aperto ai dati scientifici in quanto è un tema cruciale affinché le risorse linguistiche sviluppate in questo lavoro trovino una maggiore diffusione e abbiano un impatto effettivo sullo studio del fenomeno.

1.3 Risorse linguistiche come dati aperti

La corretta amministrazione e gestione dei dati è una componente chiave del successivo impatto di uno studio scientifico, nel campo della linguistica così come in altri campi del sapere. Per questo

nell'ambito degli studi indicati che hanno definito metodologie, strumenti e pratiche didattiche di cui il presente studio costituisce una mera applicazione.

[3] https://doi.org/10.5281/zenodo.15173712

[4] https://github.com/nicolaCirillo/termdomain

motivo, Wilkinson et al. (2016) hanno redatto i principi *Findable, Accessible, Interoperable, Reusable* (FAIR), una serie di nozioni atte a garantire la corretta amministrazione dei dati e favorirne il riuso.

Tuttavia, nell'attuale panorama delle discipline linguistiche, i principi FAIR trovano raramente applicazione e i dati raccolti dai ricercatori non vengono distribuiti con licenze aperte, non sono adeguatamente documentati oppure sono codificati in formati eterogenei. Ne consegue che molto del tempo a disposizione dei ricercatori è impiegato nella ricerca e nella conversione dei dati o nella costruzione di *dataset* simili ad altri già esistenti, che non sono stati distribuiti adeguatamente.

Perciò, per ottenere il massimo rendimento dalle risorse assegnate alla ricerca, è necessario valorizzare i dati linguistici e porre maggiore attenzione alla loro corretta amministrazione, al loro formato, alla documentazione e ai canali attraverso i quali sono distribuiti. A questo scopo, Cimiano et al. (2020) hanno verificato che i formati attualmente più diffusi per la codifica delle risorse linguistiche non rispettano pienamente il paradigma FAIR e hanno proposto quindi di codificare le risorse linguistiche tramite le tecnologie del web semantico e il paradigma dei *Linguistic Linked Open Data* (LLOD). Infatti, questo paradigma presenta numerosi vantaggi:

- Gli *Uniform Resource Identifier* (URI), garantiscono che tutti gli elementi linguistici (entrate lessicali, documenti, *token*, ecc.) siano globalmente identificabili, senza ambiguità;
- Il protocollo HTTP garantisce che ogni agente possa recuperare le informazioni e visualizzarle in diversi formati sia *machine-readable* (es. XML, JSON) che *human-readable* (es. HTML);
- Il formato *Resource Description Framework* (RDF) e il linguaggio SPARQL garantiscono un modello standard per rappresentare e interrogare i dati;
- I link tra URI creano una rete di oggetti di interesse linguistico, legati da connessioni che esprimono relazioni di diverso tipo. Ad esempio, un'entrata lessicale è connessa alle sue occorrenze in un corpus e un concetto è connesso alle entrate lessicali utilizzate per esprimerlo.

Sebbene i formati del web semantico non siano, allo stato attuale, i più utilizzati per rappresentare le risorse linguistiche, essi vedono un interesse crescente da parte della comunità scientifica, come dimostrano numerose iniziative come il progetto *Nexus Linguarum*[5] e le LDK *Conference on Language, Data and Knowledge*.

Per garantire il rispetto dei principi FAIR nel presente lavoro, abbiamo deciso di adottare i formati del web semantico e il paradigma LLOD per rappresentare le risorse linguistiche sviluppate.

1.4 Organizzazione del lavoro

Il presente lavoro è diviso in due parti principali. La prima parte è dedicata a illustrare lo sfondo teorico di riferimento e lo stato dell'arte dei diversi temi che verranno affrontati nel corso della trattazione. Nello specifico, il Capitolo 2 contiene una panoramica dei principali approcci alla terminologia intesa come scienza e, in seguito, descrive i principi teorici che sono alla base del presente studio. Il Capitolo 3 presenta l'ambito disciplinare della linguistica computazionale e le diverse tipologie di strumenti da essa sviluppati: *corpora*, *tecniche di EAT*, *strumenti di gestione terminologica* e *banche dati terminologiche*.

[5] https://nexuslinguarum.eu/ (Ultima rilevazione: 10 aprile 2025)

Successivamente, nella seconda parte del lavoro sono descritte e analizzate le risorse linguistiche realizzate. In particolare, il Capitolo 4 descrive e analizza i corpora ItaIst-DdAC_GRU e ItaIst-TermRifiuti. Nel Capitolo 5 viene illustrata e testata la tecnica di estrazione automatica di termini *Domain Concept Relatedness*. Nel Capitolo 6 è descritta la banca dati terminologica ItaIst-WasteLexicon e presentata l'analisi congiunta di ItaIst-WasteLexicon e dei corpora ItaIst-DdAC_GRU e ItaIst-TermRifiuti al fine di individuare le caratteristiche principali della terminologia istituzionale della gestione dei rifiuti urbani. Infine, nel Capitolo 7 vengono tratte le conclusioni.

Capitolo 2

Quadro teorico

Contenuti

2.1 Principali approcci alla terminologia 9
2.1.1 Teoria Classica 10
2.1.2 Approccio sociolinguistico (socioterminologia) 10
2.1.3 Approccio comunicativo 11
2.1.4 Approccio sociocognitivo 12
2.1.5 Approccio culturale 12
2.1.6 Approccio testuale 13
2.2 Sfondo teorico 14
2.2.1 Concetto di termine 14
2.2.2 Termini come unità lessicali 15
2.2.3 Le componenti del termine 17
2.2.4 Proprietà morfosintattiche del termine 18
2.2.5 Proprietà concettuali del termine 19
2.2.6 Proprietà semantico-comunicative del termine 22

2.1 Principali approcci alla terminologia

Gli studiosi sono concordi nell'individuare come padre della terminologia moderna l'ingegnere austriaco Eugene Wüster (1898 – 1977), dai cui lavori è nata la cosiddetta *Teoria Classica* della terminologia (Wüster, 1979). Le idee di Wüster hanno avuto una duplice influenza sulla terminologia odierna. Da un lato, hanno gettato le basi della programmazione linguistica e della standardizzazione terminologica, volte a favorire la comunicazione specialistica. Dall'altro, hanno acceso il dibattito scientifico. Infatti, dalle critiche mosse all'approccio wüsteriano sono nati numerosi approcci alla terminologia, che prendono le distanze dai principi della Teoria Classica e propongono un'interpretazione alternativa della disciplina (Campo, 2013). I principali sono:

- L'approccio sociolinguistico (o socioterminologia);
- L'approccio comunicativo;

- L'approccio sociocognitivo;
- L'approccio culturale;
- L'approccio testuale.

Il seguito di questa sezione illustra i principi fondamentali della Teoria Classica della terminologia e degli approcci alternativi sopra menzionati. Per una descrizione più approfondita si rimanda ai lavori di Campo (2013), Faber e L'Homme (2022) e Vezzani (2022) .

2.1.1 Teoria Classica

I principi teorici che sono alla base della Teoria Classica della terminologia sono individuabili nell'opera principale di Wüster (1979): *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie* (Introduzione alla terminologia generale e alla lessicografia terminologica).

Gli studi di Wüster hanno uno scopo pratico, orientato alla standardizzazione e pianificazione linguistica. Secondo Wüster, la terminologia è una disciplina autonoma, che si discosta dalla linguistica, in quanto non si occupa di descrivere i termini nei loro contesti d'uso ma ha un obiettivo prescrittivo: quello di favorire una comunicazione chiara e non ambigua tra esperti di settore.

Per questo motivo, la prospettiva di Wüster è strettamente onomasiologica: il centro dei suoi studi sono i concetti e le relazioni che intercorrono tra di essi. I termini invece non sono altro che etichette, le cui proprietà linguistiche non hanno rilevanza, perché lo scopo della ricerca terminologica è innanzitutto quello di fornire una chiara delimitazione dei concetti. In questo senso, Wüster ritiene che la terminologia si opponga alla linguistica: la linguistica considera la parola come unità indivisibile, composta da forma e contenuto, mentre per la Teoria Classica della terminologia, il termine e il concetto sono entità separate ed i concetti possono essere analizzati indipendentemente dai termini usati per esprimerli (Campo, 2013).

Inoltre, nella Teoria Classica, ambiguità e sinonimia sono fenomeni inaccettabili perché ostacolano la comunicazione tra esperti. Secondo Wüster, tra termini e concetti deve esserci una corrispondenza biunivoca, ovvero un termine deve esprimere un solo concetto ed un concetto deve essere espresso attraverso un solo termine. Il compito di garantire tale corrispondenza e di fornire definizioni condivise spetta alle autorità di normalizzazione linguistica, tra cui l'*International Organization for Standardization* (ISO)[1], presso la quale Wüster partecipò all'istituzione del comitato tecnico 37 che si occupa di linguaggio e terminologia (Zanola, 2018).

2.1.2 Approccio sociolinguistico (socioterminologia)

L'approccio sociolinguistico alla terminologia, anche chiamato *socioterminologia*, rappresenta il primo movimento di opposizione alla Teoria Classica. (Campo, 2013). Secondo questo approccio, la terminologia, essendo calata nei diversi contesti comunicativi dei linguaggi specialistici, è fortemente influenzata da fattori culturali, storici e geografici che devono essere accuratamente analizzati per capire a pieno i fenomeni terminologici.

Contrariamente alla Teoria Classica, che ha un forte carattere prescrittivo, l'approccio sociolinguistico è orientato alla descrizione e non considera il concetto come preesistente al termine. Al

[1] L'ISO è un'organizzazione mondiale non-governativa che si occupa della definizione di norme tecniche che definiscono le buone pratiche atte a garantire la qualità, l'affidabilità e l'efficienza dei processi aziendali e di altro tipo.

contrario, il linguista francese François Gaudin (1993, 2003) sostiene che concetto e termine si formino insieme, all'interno dei discorsi dove sono continuamente sottoposti a tensioni e cambiamenti provocati da fattori sociali e storici.

In altre parole, per la socioterminologia, i concetti non sono entità statiche, bensì si evolvono nel tempo, con l'evolversi delle discipline di cui essi fanno parte. Per questo motivo, il metodo di studio della socioterminologia è di carattere semasiologico, ovvero parte dall'analisi dei termini per arrivare alla definizione dei concetti.

Anche il principio di biunivocità è messo in dubbio. I socioterminologi sostengono che, essendo sinonimia e polisemia caratteristiche proprie del linguaggio, esse caratterizzano inevitabilmente anche i linguaggi specialistici e non possono perciò essere ignorate.

Inoltre, la socioterminologia non condivide gli obiettivi di pianificazione linguistica come intesi da Wüster. Sostiene questa posizione anche il linguista delle scienze traduttive Yves Gambier (1998), affermando che l'ossessione di creare un linguaggio non ambiguo, attraverso le norme imposte dagli organismi internazionali di standardizzazione, crea un clima dittatoriale e ostacola l'evoluzione e lo sviluppo della conoscenza che dovrebbe invece emergere naturalmente dalle comunità di discorso.

2.1.3 Approccio comunicativo

L'approccio comunicativo alla terminologia nasce dagli studi della linguista catalana Maria Teresa Cabré Castellví (1998) e si oppone alla visione della terminologia come disciplina autonoma e indipendente dalla linguistica. Al contrario, l'approccio comunicativo fa propria la prospettiva linguistica e concentra l'attenzione sul rapporto tra variazione terminologica e contesti comunicativi. In altre parole, non tratta la terminologia come un'attività denominativa, che ha lo scopo di identificare ed etichettare i concetti, ma come un'attività comunicativa, dove il termine, che nell'approccio comunicativo prende il nome di *unità terminologica*, rappresenta un determinato concetto in un determinato contesto d'uso.

Infatti, Cabré Castellví (1998) non considera i termini come entità indipendenti dalle parole della lingua comune, bensì li considera parte integrante del lessico e sottolinea come conoscenze generali e specialistiche siano interconnesse nella mente dei parlanti. Infatti, nella prospettiva dell'approccio comunicativo, le unità lessicali di una lingua, se prese fuori da un determinato contesto, non sono né termini né parole. L'unità terminologica emerge quando un'unità lessicale è calata in un contesto comunicativo dove possiede una specifica funzione d'uso.

Per questo motivo, l'approccio comunicativo alla terminologia studia i termini *in vivo*, nella produzione linguistica spontanea (scritta e orale), contrapponendosi allo studio *in vitro* tipico della standardizzazione e pianificazione linguistica. Inoltre, l'approccio comunicativo contempla l'esistenza di fenomeni quali polisemia e sinonimia, riconoscendo che i termini possono entrare a far parte del linguaggio comune e che le parole comuni possono terminologizzarsi, assumendo significati specialistici (Cabré Castellví, 1998).

Inoltre, Cabré Castellví (1998) ritiene che un dato termine non appartenga esclusivamente a un unico dominio di conoscenza. Sostiene invece che i termini, sebbene nascano nell'ambito di un determinato dominio di conoscenza, vengano poi usati in più domini o anche nella lingua comune, a causa del *transfert* terminologico da un dominio all'altro o da un linguaggio specialistico verso la lingua comune e viceversa.

2.1.4 Approccio sociocognitivo

L'approccio sociocognitivo alla terminologia nasce dagli studi della linguista belga Rita Temmerman (2000) che, nel descrivere i termini delle scienze naturali, trova i principi della Teoria Classica inadeguati. Di conseguenza, adotta una prospettiva mutuata dalla semantica cognitiva, secondo la quale la conoscenza linguistica non sarebbe diversa dagli altri tipi di conoscenza. In questa prospettiva, il significato di parole e termini esiste nella mente dei parlanti come frutto di un processo cognitivo e può essere compreso solo in riferimento ad un insieme di esperienze, valori o pratiche che formano una struttura organizzata di conoscenza.

Inoltre, Temmerman (2000) respinge la nozione classica di concetto definito in termini di estensione ed intensione a favore di una definizione basata sulla struttura prototipica e la somiglianza di famiglia, per la quale conia il termine *unità di comprensione*. Le unità di comprensione sono interpretate alla luce di strutture cognitive più ampie, insiemi strutturati di concetti, dove per comprendere un determinato concetto, è necessario aver presente lo schema concettuale di cui esso fa parte. Tali schemi sono i *frame* di Charles J. Fillmore (1985) ed i *modelli cognitivi idealizzati* di George Lakoff (1987).

Un altro aspetto chiave dell'approccio sociocognitivo è la rilevanza dello studio diacronico, usato specialmente come mezzo per spiegare la polisemia. Secondo Temmerman (2000), quando viene scoperto un nuovo fenomeno, si tende ad assegnargli un termine univoco, con un solo significato. Poi, man mano che la conoscenza del fenomeno si estende e viene approfondita, questo termine si arricchisce di nuovi significati e tende alla polisemia. In questo senso, la polisemia è indice e conseguenza dello sviluppo di nuova conoscenza.

Infine, l'approccio sociocognitivo, oltre alla terminologia in sé, intende studiare anche i fattori tecnici e sociologici che favoriscono o inibiscono particolari fenomeni di creazione e comunicazione della conoscenza. Le stesse dinamiche di potere alla base della pianificazione linguistica, secondo Temmerman (2000), non dovrebbero essere accettate aprioristicamente ma attentamente analizzate dal punto di vista storico e sociologico.

2.1.5 Approccio culturale

L'approccio culturale nasce dagli studi di Diki-Kidiri nello specifico contesto delle lingue africane, con l'obiettivo di favorire l'ingresso di nuovi saperi nelle comunità che parlano queste lingue. Per raggiungere questo obiettivo, Diki-Kidiri (2022) intende impiegare la terminologia come strumento. Infatti, egli afferma che queste comunità possono fare proprie le conoscenze prodotte da altre culture ed estendere il loro bagaglio conoscitivo attraverso lo sviluppo di nuove terminologie scientifiche.

Secondo l'approccio culturale, il modo in cui gli individui e le comunità percepiscono e comprendono il mondo è fortemente influenzato dalla cultura, definita come l'insieme di esperienze, conoscenze e attività che contribuiscono a formare l'identità di una persona o di una comunità in un determinato luogo e in una determinata epoca.

Ecco perché Diki-Kidiri (2022) ritiene fondamentale includere la dimensione culturale nella ricerca terminologica, dimensione che è totalmente assente nella Teoria Classica di stampo wüsteriano, che perciò non è equipaggiata per far fronte allo scambio di conoscenze tra comunità appartenenti a culture diverse.

Anche la tradizionale dicotomia termine/concetto è inadatta a catturare l'aspetto culturale dei fenomeni terminologici. Per questo motivo, l'approccio culturale la abbandona a favore di una scomposizione del termine in tre elementi: il *significante*, il *percetto* (o significato) e il *concetto*.

Il significante appartiene alla dimensione grafica e fonetica del termine, esso è la sua rappresentazione linguistica come viene letta o udita dai parlanti. Il percetto e il concetto, invece, appartengono entrambi alla dimensione semantica. Nello specifico, il concetto fa riferimento alla classi di oggetti a cui il termine si riferisce ed è essenzialmente oggettivo ed universale. Il percetto, invece, fa riferimento alla costruzione del significato. Quest'ultima non è universale ma strettamente legata a come una data cultura percepisce il concetto espresso dal termine.

Infine, l'approccio culturale definisce due fasi del lavoro terminologico: la ricerca terminologica e lo sviluppo di terminologie. La prima richiede il lavoro di linguisti e specialisti di dominio che parlano lingue diverse e provengono da diverse culture mentre la seconda prevede delle fasi molto precise. Esse consistono in:

1. Definire lo scenario sociale e gli obiettivi di ricerca;
2. Scegliere un insieme di termini da un sistema concettuale;
3. Studiare il loro uso e la loro evoluzione;
4. Individuare i loro equivalenti nella lingua target;
5. Formare neologismi (o ricorrere a prestiti) per i termini che non hanno equivalenti nella lingua target;
6. Valutare gli equivalenti insieme a parlanti nativi ed esperti di dominio;
7. Standardizzare la terminologia, selezionando (dove ciò sia possibile) un solo termine per ogni concetto;
8. Diffondere i termini standardizzati.

2.1.6 Approccio testuale

Negli ultimi decenni, l'utilizzo di strumenti digitali ha fortemente modificato le pratiche terminografiche e lo studio della terminologia. Corpora di grandi dimensioni, compilatori di concordanze, algoritmi di estrazione di termini e banche dati terminologiche sono tutti strumenti a cui i terminologi odierni non possono rinunciare. In questo contesto, nasce l'approccio testuale alla terminologia, un approccio strettamente legato alla linguistica dei corpora e al trattamento automatico del linguaggio.

L'approccio testuale si oppone fermamente alla prospettiva onomasiologica della Teoria Classica, che parte dall'analisi del concetto per arrivare al termine. Piuttosto, l'approccio testuale adotta una prospettiva semasiologica, dove lo studio dei termini parte necessariamente dal testo che rappresenta per la terminologia testuale una preziosa fonte di evidenza empirica, nella quale i termini sono osservabili in contesti comunicativi reali (Bourigault & Slodzian, 1999; L'Homme, 2006). In questo senso, l'analisi terminologica presenta numerose similitudini con le pratiche lessicografiche.

Inoltre, l'approccio testuale rifiuta di considerare i termini come entità esterne alla lingua. Ritiene invece che i termini appartengano al lessico e che quindi la terminologia abbia il compito di analizzarne gli aspetti lessicali, sintattici, semantici e pragmatici.

Per questo motivo, tale analisi non può essere demandata esclusivamente ad esperti di dominio. Deve invece avvalersi della figura del linguista-terminologo, che ha il compito di costruire risorse terminologiche sulla base delle occorrenze attestate, guidato sia dalla specificità del corpus sia dagli obiettivi dell'applicazione pratica. Inoltre, il linguista-terminologo collabora con gli esperti di dominio per validare descrizioni e scelte terminologiche (Bourigault & Slodzian, 1999).

Per compiere le proprie ricerche, il linguista-terminologo impiega dei corpora specialistici, costruiti in modo tale da rappresentare adeguatamente il fenomeno oggetto di indagine. Nello specifico, un corpus specialistico deve avere le seguenti caratteristiche (Bowker & Pearson, 2002):

- I testi devono rappresentare situazioni comunicative reali;
- I testi devono essere in formato elettronico per poter essere analizzati tramite strumenti di trattamento automatico del linguaggio;
- Il corpus deve avere una dimensione adeguata allo scopo dello studio;
- Il testi devono essere scelti secondo specifici criteri che ne garantiscano la rappresentatività.

Una volta che il corpus è stato realizzato, esso può essere analizzato grazie all'ausilio di strumenti di estrazione automatica di termini e di relazioni tra termini, che devono poi essere validati da esperti di dominio.

Con l'ausilio di questi strumenti, la terminologia testuale ha realizzato applicazioni terminologiche che soddisfano specifiche necessità sociali: vocabolari standard per l'indicizzazione di documenti, terminologie per i software di traduzione, ontologie, basi di conoscenza terminologiche, ecc. (Campo, 2013). Il Capitolo 3 presenta una rassegna di questi strumenti.

2.2 Sfondo teorico

Il presente lavoro si colloca nel quadro degli studi terminologici applicati ai processi di comunicazione pubblica e istituzionale (Vellutino, 2014, 2018, 2022; Vellutino et al., 2012, 2016). Tali studi hanno delineato un approccio teorico-metodologico che rielabora in modo innovativo e coerente gli approcci descritti in letteratura, in particolare, l'approccio comunicativo e quello testuale.

Pertanto, seguendo tale approccio, come emergerà dal seguito, il presente lavoro focalizza il rapporto tra variazione terminologica e contesti comunicativi, mentre dall'approccio testuale è tratta l'attenzione ai dati empirici, ai corpora e alle banche dati terminologiche, l'uso di strumenti digitali e lo sviluppo di risorse terminologiche.

2.2.1 Concetto di termine

Definire il concetto di *termine* non è un compito semplice. La difficoltà principale risiede nel fatto che i termini hanno le stesse caratteristiche formali delle altre parole di una lingua, per cui risulta problematico tracciare una distinzione netta tra termini e parole comuni.

Secondo la Teoria Classica della terminologia, un termine è un simbolo convenzionale che designa un concetto, nell'ambito di un determinato dominio di conoscenza. Una definizione simile è contenuta nella norma ISO 1087:2019, dove *termine* è definito come "designazione che rappresenta un concetto generale per mezzo della lingua." Risulta chiaro, quindi, che i termini hanno almeno due dimensioni: una linguistica ed una concettuale, sono unità linguistiche che denotano un concetto (Zanola, 2018).

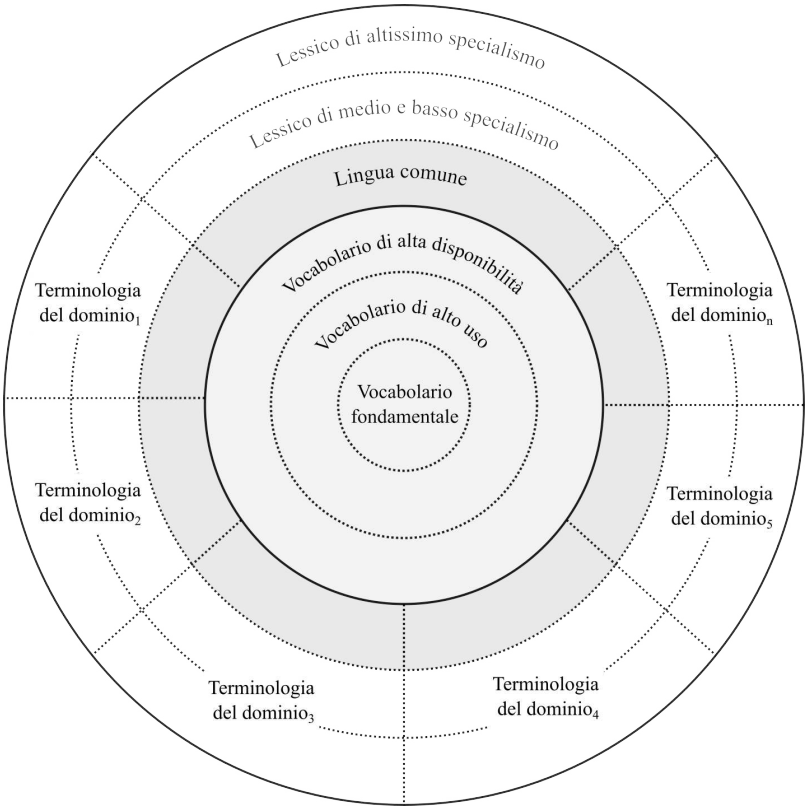


Figura 2.1: Organizzazione del lessico. *Immagine adattata da Gualdo e Telve (2021).*

Il limite di questa definizione è che essa non permette di distinguere tra termini e parole comuni. Nella pratica, se provassimo ad applicare questa definizione alle unità lessicali di un testo specialistico, ci accorgeremmo che essa includerebbe tutte le parole semanticamente piene.

Questo avviene perché, di fatto, i termini appartengono allo stesso sistema linguistico delle parole (Cabré Castellví, 1999). Anzi, potremmo dire che i termini sono unità lessicali a tutti gli effetti, sebbene appartengano al lessico dei linguaggi speciali. Questi ultimi, infatti, non sono codici a sé stanti, bensì sono varietà di una determinata lingua naturale (Auger & L'Homme, 1994/1995).

2.2.2 Termini come unità lessicali

Il lessico di una lingua è quindi formato da un nucleo di parole fondamentali molto frequenti e note a tutti i parlanti, da parole meno comuni, note solo a parlanti colti, e da termini specialistici, noti soltanto agli specialisti di una determinata disciplina (Figura 2.1). In particolare, Gualdo e Telve (2021), riprendendo gli studi di De Mauro e Chiari (2016) e di Serianni (2007), individuano per l'italiano quattro livelli lessicali: il vocabolario di base, il lessico comune, il lessico di medio-alto specialismo e il lessico di altissimo specialismo.

Il vocabolario di base rappresenta il nucleo della lingua e contiene circa 6.000 parole ad altissima frequenza (De Mauro & Chiari, 2016). Segue il lessico comune, che comprende tra le 45.000 e 50.000 parole (De Mauro & Chiari, 2016) e costituisce una zona di transizione tra il lessico di base e quello specialistico. Infine, troviamo i lessici di medio-alto specialismo e altissimo specialismo (Serianni,

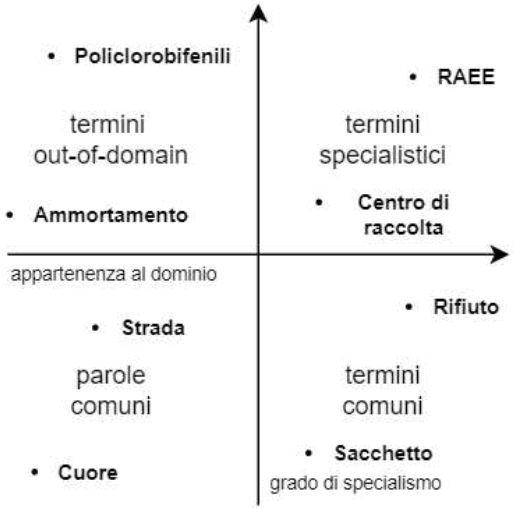


Figura 2.2: Rappresentazione dei termini sulla base del grado di specialismo e dell'appartenenza al dominio. *Immagine adattata da Rigouts Terryn et al. (2020)*

2007) che contengono i termini di tutti i domini di conoscenza, la cui dimensione è difficile stimare. Si noti inoltre che i confini tra queste aree sono molto sfumati e che, spesso, gli strati più vicini tendono a influenzarsi a vicenda.

Dunque, per distinguere i termini dalle altre parole, è utile prendere in considerazione due dimensioni: lo specialismo e l'appartenenza al dominio (Rigouts Terryn et al., 2020), come illustrato nello schema in Figura 2.2. Lo specialismo, come in Serianni (2007) e Vellutino (2018), fa riferimento ai parlanti che conoscono e utilizzano un dato termine. Un termine di alto specialismo è noto a una ristretta cerchia di esperti di settore, mentre un termine di basso specialismo è conosciuto da un gruppo ampio ed eterogeneo di parlanti. Il grado di appartenenza al dominio, invece, indica la centralità di un termine rispetto a un determinato dominio di conoscenza.

I termini di alto specialismo si distinguono facilmente dalle parole comuni. Ad esempio, è intuitivo considerare come unità terminologiche espressioni specialistiche come *compostiera elettromeccanica* e *rete da circuizione*. Al contrario, la situazione dei termini di medio e basso specialismo è più complessa perché questi ultimi tendono a fondersi con gli strati più esterni della lingua comune. Prendiamo ad esempio le parole *lenza* e *discarica*. Qui il grado di appartenenza al dominio riveste un ruolo fondamentale. Infatti, se facciamo riferimento al dominio della pesca, considereremo *lenza* un termine e *discarica* no. Invece, se il nostro riferimento fosse il dominio dell'ambiente, avverrebbe esattamente l'opposto: *discarica* sarebbe un termine mentre *lenza* no.

A questo proposito, nel loro schema di annotazione terminologica, Rigouts Terryn et al. (2020) distinguono tre tipologie di termini: i termini specialistici, i termini comuni e i termini *out-of-domain*.

I termini specialistici sono unità lessicali specializzate e appartenenti al dominio considerato in un determinato studio; i termini comuni sono unità lessicali che appartengono al dominio ma sono note a numerosi parlanti e i termini *out-of-domain* sono unità lessicali che sono note solo agli specialisti ma non appartengono al dominio considerato. Infine, le unità lessicali che non sono né specialistiche né appartenenti al dominio non sono da considerarsi termini ma parole comuni.

Inoltre, le differenze tra i termini e le altre parole emergono con più chiarezza quando l'oggetto di analisi si sposta dalla singola unità terminologica ad un'intera terminologia, intesa come l'insieme di tutti i termini che appartengono ad un determinato dominio di conoscenza o linguaggio specialistico. Infatti, le unità lessicali di una terminologia sono molto legate tra di loro sia a livello linguistico che concettuale e tendono a formare strutture specifiche, cosa che avviene con meno frequenza tra termini afferenti a discipline diverse e lontane (Cabré Castellví, 1999).

Anche i meccanismi di formazione dei termini sono gli stessi che formano le altre parole: prefissazione, suffissazione, espansione semantica, ecc. Ciò che cambia è la produttività di tali meccanismi. Ad esempio, il morfema *cardio* è molto più produttivo nel linguaggio speciale della medicina che in quello comune.

Alla luce di queste considerazioni proveremo a fornire una definizione di termine in linea con lo scopo del presente lavoro.

> **Termine:** Unità lessicale calata in un dato contesto comunicativo che, in relazione a tale contesto, presenta almeno una delle seguenti caratteristiche:
>
> - è raramente utilizzata nella comunicazione ordinaria tra non esperti;
> - esprime un concetto rilevante rispetto ai domini di conoscenza che sono centrali all'interno di tale contesto comunicativo.

2.2.3 Le componenti del termine

Abbiamo visto che i termini hanno una duplice natura: linguistica e concettuale. La natura linguistica comprende sia il significante che il significato del termine, che è fortemente influenzato dal sistema-lingua a cui tale termine appartiene . Invece, la natura concettuale risiede nel concetto denotato e nelle relazioni che esso intrattiene con altri concetti. A queste due componenti, quella linguistica e quella concettuale, Cabré Castellví (2003) propone di aggiungere una terza componente, la componente comunicativa, che si riferisce ai contesti comunicativi in cui il termine è impiegato.

Ai fini di questa trattazione, adotteremo queste tre componenti ma in una diversa configurazione. In particolare, la componente linguistica è stata scorporata e messa in relazione con quella comunicativa, dando luogo alla componente morfosintattica e alla componente semantico-comunicativa. La componente concettuale resta invece invariata. Di seguito, è descritto come queste tre componenti siano da intendersi ai fini della presente trattazione.

Componente morfosintattica. La componente morfosintattica di un termine fa riferimento alla forma in cui tale termine è pronunciato/udito e scritto/letto. In altre parole, la componente morfosintattica di un termine corrisponde alla sua forma di superficie, al suo significante. Inoltre, la componente morfosintattica comprende anche le proprietà grammaticali di tale forma.

Componente concettuale. La componente concettuale di un termine include la porzione denotativa del significato di tale termine, che delimita la classe di oggetti che il termine sta ad indicare, e la sua intensione, ovvero l'insieme di caratteristiche e proprietà possedute da tale classe di oggetti.

Componente semantico-comunicativa. La componente semantico-comunicativa di un termine comprende la porzione connotativa di significato di tale termine, intesa come l'insieme delle

caratteristiche valutative che evoca e che non sono necessariamente definitorie della classe di oggetti denotata. La componente comunicativa comprende anche il valore pragmatico-situazionale del termine con particolare riguardo ai contesti comunicativi in cui viene impiegato.

Nei paragrafi seguenti elencheremo e analizzeremo le caratteristiche principali di ognuna di queste tre componenti dei termini.

2.2.4 Proprietà morfosintattiche del termine

Come già accennato in precedenza, i termini sono di fatto unità lessicali. Per questo motivo, essi ereditano tutte le proprietà che caratterizzano gli elementi del lessico (Cabré Castellví, 1998; Vellutino, 2018), prima fra tutte la classe lessicale (o parte del discorso) che determina anche le proprietà grammaticali. I termini hanno inoltre una strutturazione interna complessa che può essere oggetto di analisi morfologica, per quanto riguarda i termini monorematici, o sintattica, per quanto riguarda i termini polirematici.

Parte del discorso. Generalmente, i termini appartengono alle classi lessicali aperte: sono principalmente nomi (es. *rifiuti urbani*, *conferimento*, *trattamento termico*) ma possono essere anche verbi (es. *conferire*, *compostare*, *cernire*), aggettivi (es. *indifferenziato*, *differenziato*, *conferibile*) e avverbi (es. *sostenibilmente*, *gerarchicamente*).

La terminologia di un determinato dominio contiene spesso alcuni termini che presentano lo stesso morfema lessicale ma appartengono a categorie lessicali diverse, in funzione del diverso morfema derivazionale. Ad esempio, i termini *conferire*, *conferimento* e *conferibile* sono tre termini appartenenti al dominio della gestione dei rifiuti urbani. Tutti e tre hanno come base il morfema lessicale *confer-* ma il primo è un verbo, il secondo un nome e il terzo un aggettivo.

Fondamentali sono anche le proprietà grammaticali dei termini: il genere e il numero per nomi e aggettivi; il modo, il tempo, la persona e il numero per i verbi.

Struttura morfosintattica. Altra caratteristica fondamentale di un termine è la sua struttura morfosintattica. In particolare, la struttura morfosintattica di un termine può appartenere ad una delle seguenti categorie.

- monorematica: il termine è composto da una sola parola grafica, senza spazi (es. *coinceneri-mento*, *indifferenziato*, *biodigestore*);
- polirematica: il termine è composto da più parole grafiche, separate da spazi (es. *tal quale*, *incenerimento mediante ossidazione*, *frazione organica del rifiuto solido urbano*);
- inizialismo o acronimo: il termine è composto dalle iniziali di un'unità polirematica (es. *RUP – Rifiuti Urbani Pericolosi*; *FORSU – Frazione Organica del Rifiuto Solido Urbano*)
- abbreviazione: il termine è una forma abbreviata di un'unità monorematica o polirematica (es. *Art. – Articolo*; *D.lgs – Decreto Legislativo*)
- struttura mista: il termine è un'unità polirematica che contiene al suo interno inizialismi e/o abbreviazioni (es. *Vi.Vi.Fir – Vidimazione Virtuale del Formulario di Identificazione dei Rifiuti*; *ciclo integrato dei RU – ciclo integrato dei Rifiuti Urbani*).

In una terminologia, capita spesso che uno stesso concetto sia espresso attraverso più forme. Ad esempio, il termine *raccolta differenziata* è un'unità polirematica che può essere acronimizzata in *RD* e che possiede anche una variante monorematica a basso grado di specialismo, *la differenziata*.

Struttura interna dei termini. Oggetto di analisi terminologica è anche la struttura interna dei termini. I termini polirematici possono essere composti a loro volta da altri termini che chiameremo sotto-termini. Ad esempio, il termine *rifiuti urbani misti* contiene al suo interno il sotto-termine *rifiuti urbani*. O ancora il termine *gestione integrata dei rifiuti urbani* contiene il sotto-termine *gestione dei rifiuti urbani* che, in questo caso, è discontinuo. Ovvero non appare nel testo come una sequenza continua di parole ma è separato da materiale linguistico estraneo al termine stesso.

Anche la struttura morfologica dei termini monorematici è di interesse per la terminologia. Ad esempio, possiamo notare come numerosi termini del dominio della gestione dei rifiuti urbani contengano i prefissi *bio-* ed *eco-* (es. *biodigestore*; *biodegradabile*; *ecocentro*; *ecocompattatore*)

2.2.5 Proprietà concettuali del termine

Le proprietà concettuali di un termine caratterizzano il concetto espresso da tale termine e definiscono le relazioni che esso intrattiene con altri concetti.

Definizione. Di fondamentale importanza è la definizione del concetto. Essa serve a determinare quali oggetti corrispondono a tale concetto e sono quindi referenti del termine. Secondo la norma ISO 1087:2019 (2019), una definizione può essere di due tipi: intensionale o estensionale.

Una definizione intensionale elenca le caratteristiche necessarie e sufficienti che un oggetto deve possedere per corrispondere al concetto (l'intensione del concetto). Di questo tipo è la definizione di *rifiuto* contenuta nell'articolo 3 della Direttiva 2008/98/CE del Parlamento europeo e del Consiglio, del 19 novembre 2008, relativa ai rifiuti e che abroga alcune direttive.

> **rifiuto:** qualsiasi sostanza od oggetto di cui il detentore si disfi o abbia l'intenzione o l'obbligo di disfarsi.

La definizione estensionale, invece, non indica delle proprietà ma elenca tutti gli oggetti che ricadono sotto il concetto definito (l'estensione del concetto). Di questo tipo è la definizione di *rifiuti organici* contenuta nell'articolo 3 della Direttiva 2008/98/CE.

> **rifiuti organici:** Rifiuti biodegradabili di giardini e parchi, rifiuti alimentari e di cucina prodotti da nuclei domestici, ristoranti, servizi di ristorazione e punti vendita al dettaglio e rifiuti simili prodotti dagli impianti dell'industria alimentare.

Non sempre è possibile definire un concetto in termini di intensione ed estensione in quanto alcuni concetti sono impossibili da formalizzare in termini di proprietà necessarie e sufficienti. Per questo motivo, Temmerman (2000) propone di definire questi concetti come strutturati prototipicamente, elencandone le caratteristiche più rilevanti e tenendo conto della loro evoluzione storica.

Relazioni semantiche. Altra proprietà fondamentale di un concetto sono le relazioni che intercorrono tra di esso e altri concetti. Di fatto tali relazioni contribuiscono a definire e caratterizzare il concetto stesso. Le relazioni più utilizzate sono quelle tassonomiche, che permettono di organizzare

i concetti su base gerarchica, la più importante delle quali è la relazione *genere/specie*. La relazione *genere/specie* indica che un concetto specifico è una forma particolare di un concetto più generale. Ad esempio, i *Rifiuti da Apparecchiature Elettriche ed Elettroniche* (RAEE) sono un tipo particolare di *rifiuti*. In linguistica, questa relazione semantica corrisponde all'iponimia mentre il suo opposto è l'iperonimia (es. *RAEE* è iponimo di *rifiuto* e *rifiuto* è iperonimo di *RAEE*).

La relazione *parte/tutto*, invece, è una relazione semantica in cui un concetto è una componente di un altro concetto. Ad esempio, il *cassone scarrabile* è una componente del *centro di raccolta*. In linguistica, la relazione *parte/tutto* corrisponde alla meronimia ed il suo opposto è l'olonimia (es. *cassone scarrabile* è meronimo di *centro di raccolta* e *centro di raccolta* è olonimo di *cassone scarrabile*).

Dominio di conoscenza. Il dominio di conoscenza si riferisce a un'area del sapere o ad un campo specifico di studio o di interesse. È nel contesto dei domini di conoscenza che i concetti e i termini che li esprimono vengono compresi e utilizzati. Esempi di domini di conoscenza includono la matematica, la fisica, la medicina e l'informatica.

I domini di conoscenza, però, non sono da intendersi come compartimenti stagni ma, al contrario, hanno confini molto sfumati e spesso si intersecano, influenzandosi a vicenda (Boulanger, 1995; Temmerman, 2000). In particolare Vellutino (2018, p. 51) afferma che

> A livello terminologico, lessicale e morfosintattico, l'italiano istituzionale è il risultato della combinazione della lingua comune con i linguaggi istituzionali speciali dominanti del diritto, dell'amministrazione, dell'economia e della finanza pubblica, ai quali si uniscono le terminologie specialistiche dei domini di conoscenza degli argomenti dei settori delle specifiche istituzioni.

Per questo motivo, ai fini di questa trattazione, riteniamo utile rappresentare i domini di conoscenza come una rete interconnessa di saperi, secondo un modello gerarchico che chiameremo modello degli strati terminologici (Cirillo, 2024).

Secondo Drouin et al. (2018a), la terminologia di domini particolarmente vasti, come ad esempio quello relativo all'ambiente, che comprende una varietà di temi diversi (cambiamento climatico, energie rinnovabili, inquinamento, ecc.), si divide in due strati che chiameremo lessico generico di dominio e lessico tematico specifico.

Il lessico generico di dominio include quei termini che sono presenti in tutti i testi relativi all'ambiente, indipendentemente dal tema specifico trattato. Ad esempio, *biologo*, *ecosistema* e *sostenibile*. Invece, il lessico tematico specifico contiene quei termini che sono specifici di un determinato tema. Ad esempio, termini come *pala eolica* e *fotovoltaico* appartengono al lessico tematico specifico delle energie rinnovabili mentre *riscaldamento globale* ed *effetto serra* appartengono a quello del cambiamento climatico.

Astraendo ulteriormente questo principio, possiamo affermare che un dominio può essere separato in sottodomini più specifici che a loro volta sono scomponibili in ulteriori sottodomini ancora più specializzati, fino a formare un complesso modello gerarchico. Chiameremo strato terminologico (ST) la terminologia di un determinato dominio o sotto-dominio, indipendentemente dalla sua posizione nella gerarchia (Cirillo, 2024).

Gli strati terminologici hanno le seguenti proprietà:

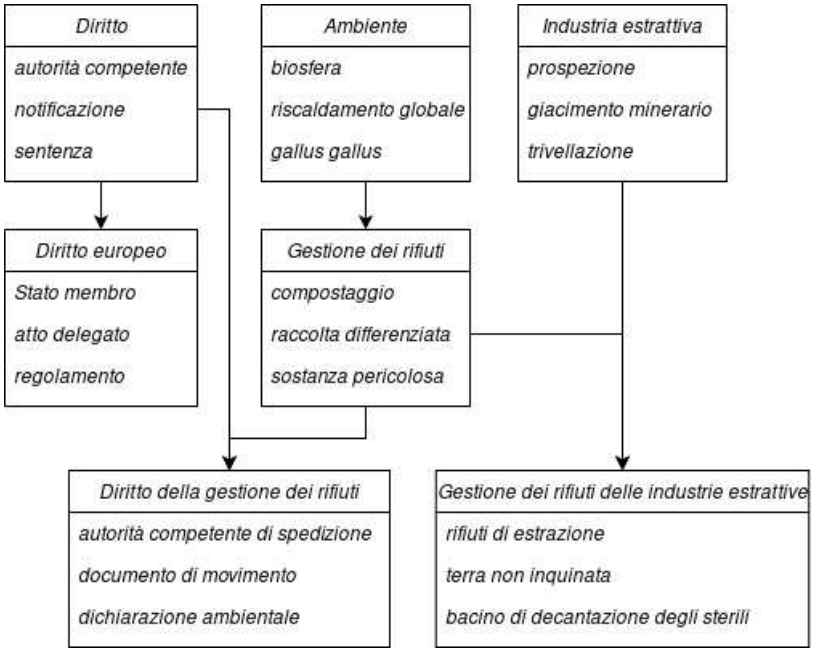


Figura 2.3: Strati terminologici. *Immagine adattata da Cirillo (2024).*

- Da un ST più generico possono discendere degli ST più specifici; Ad esempio, l'ST "gestione dei rifiuti" discende dall'ST "ambiente".
- Un ST può discendere da più ST generici; Ad esempio, l'ST "normativa sulla gestione dei rifiuti" discende dall'ST "gestione dei rifiuti" e dall'ST "diritto".
- Un ST include anche la terminologia di tutti gli ST che discendono da esso (direttamente o indirettamente); Quindi, l'ST "ambiente" includerà anche i termini dell'ST "gestione dei rifiuti" che a sua volta conterrà anche i termini dell'ST "normativa della gestione dei rifiuti".

Per comprendere meglio il concetto di ST analizziamo un estratto della "Direttiva 2006/21/CE del Parlamento europeo e del Consiglio, del 15 marzo 2006, relativa alla gestione dei rifiuti delle industrie estrattive e che modifica la direttiva 2004/35/CE"

> Prima dell'avvio di qualunque operazione che comporti l'accumulo o il deposito dei **[1] rifiuti di estrazione** in una **[2] struttura di deposito dei rifiuti**, **[3] l'autorità competente** chiede una **[4] garanzia finanziaria** (per esempio sotto forma di cauzione, compresi i fondi di garanzia mutualistici finanziati dall'industria) o altro strumento equivalente, secondo le procedure che saranno decise dagli **[5] Stati membri**, [...]

I termini [2], [3] e [4] sono facilmente inquadrabili: appartengono rispettivamente agli ST "gestione dei rifiuti", "diritto" e "finanza". Invece, il termine [5], sebbene sia anche esso un termine legale, è più specifico del termine [2] ed appartiene quindi all'ST "diritto dell'Unione Europea" che discende dall'ST "diritto". Invece, per assegnare il termine [1] all'ST corretto dobbiamo esaminare le parole che lo compongono. Esse sono a loro volta termini: il termine rifiuto appartiene all'ST "gestione dei rifiuti" mentre il termine estrazione appartiene all'ST "industrie estrattive". Ne consegue che il

termine [1] appartiene all'ST "gestione dei rifiuti delle industrie estrattive" che discende sia dall'ST "gestione dei rifiuti" che dall'ST "industrie estrattive".

2.2.6 Proprietà semantico-comunicative del termine

Come già accennato in precedenza molto spesso lo stesso concetto è espresso da più termini, chiamati varianti terminologiche. Ciò che differenzia un termine dall'altro è la connotazione: una variante può essere più o meno adatta a seconda del contesto comunicativo in cui viene utilizzata (Vellutino, 2015, 2018, 2022).

Ad esempio, il termine *centro di raccolta comunale* definito come "area comunale adibita alla raccolta differenziata dei rifiuti urbani" presenta numerose varianti. Alcune di esse sono semplici permutazioni come *centro comunale di raccolta*. Altre, presentano l'aggiunta di materiale lessicale atto a rendere il termine più trasparente e meno ambiguo, come *centro comunale di raccolta rifiuti*. Altre varianti ancora sono acronimi che servono ad abbreviare il testo e sono usati spesso in database e tabelle (es. *CRC – Centro di Raccolta Comunale*; e *CCR – Centro Comunale di Raccolta*).

Infine, abbiamo delle varianti fortemente connotate: *isola ecologica comunale* ed *ecocentro comunale*. Queste varianti a basso grado di specialismo appartengono a un registro più informale e, tramite l'aggettivo *ecologica* nel primo caso e il prefisso *eco-* nel secondo, veicolano il valore positivo del rispetto per l'ambiente. Per questo motivo, sono più utilizzate nella comunicazione tra esperti e non esperti.

Capitolo 3

Terminologia computazionale

Contenuti

3.1 Introduzione 23
3.2 Corpora specialistici e di dominio 24
3.2.1 Corpora in terminologia 25
3.2.2 Corpora specialistici primari 26
3.2.3 Corpora specialistici secondari 27
3.2.4 Formati di rappresentazione 28
3.3 Estrazione automatica di termini 30
3.3.1 Approcci tradizionali 30
3.3.2 Approcci basati su *deep learning* e *prompting* 31
3.3.3 Tecniche per isolare uno strato terminologico 32
3.4 Gestione e pubblicazione delle risorse terminologiche 33
3.4.1 Formati di rappresentazione 33
3.4.2 Principali banche dati terminologiche istituzionali 37

3.1 Introduzione

Nella terminologia contemporanea, le applicazioni informatiche sono strumenti indispensabili per produrre, analizzare e pubblicare dati terminologici. Tra di esse troviamo i corpora specialistici e di dominio, ovvero raccolte di testi specialistici appositamente selezionati e resi disponibili in formato elettronico con vari livelli di annotazione. Questi corpora offrono al terminologo una base empirica per analizzare e ricercare i termini nei loro contesti d'uso. Abbiamo poi i software di gestione terminologica che consentono di consultare, sviluppare, aggiornare e pubblicare banche dati terminologiche monolingue e multilingue in un ambiente di sviluppo collaborativo. Anche gli algoritmi di Estrazione Automatica di Termini (EAT) giocano un ruolo cruciale, permettendo di generare automaticamente liste di termini a partire da un corpora specialistici. Inoltre, l'avvento e la diffusione dell'intelligenza artificiale generativa ha aperto la strada allo sviluppo di algoritmi per la creazione automatica di definizioni, atti a facilitare ulteriormente le pratiche lessicografiche.

La disciplina che si occupa dello sviluppo e della valutazione di questi strumenti si chiama terminologia computazionale ed è una branca del Trattamento Automatico del Linguaggio (TAL). La terminologia computazionale coinvolge informatici, linguisti e terminologi e ha l'obiettivo di sviluppare risorse terminologiche e sistemi di trattamento automatico dei dati terminologici (corpora specialistici; tesauri, banche dati terminologiche, ontologie, ecc.)

3.2 Corpora specialistici e di dominio

Un corpus è una raccolta di testi, più o meno ampia, che fornisce dati empirici per lo studio di fenomeni linguistici. I corpora sono oggetto di studio della linguistica dei corpora, una branca della linguistica che non è caratterizzata da specifici obiettivi di studio, come ad esempio la sociolinguistica o la psicolinguistica ma è di carattere strettamente metodologico. Biber e Reppen (2015, p. 1) affermano che

> La caratteristica distintiva della linguistica dei corpora è l'affermazione che sia effettivamente possibile "rappresentare" un campo d'uso del linguaggio con un corpus di testi, e che sia possibile descrivere empiricamente schemi d'uso linguistici attraverso l'analisi di quel corpus. Ogni domanda di ricerca relativa all'uso e alla variazione linguistica può essere affrontata da questa prospettiva metodologica.

Pertanto, per poter generalizzare i risultati ottenuti dall'analisi di un corpus, quest'ultimo deve essere rappresentativo della popolazione di testi da cui è tratto. Tuttavia, costruire un corpus è un compito complesso e ottenere un corpus che sia totalmente rappresentativo del fenomeno oggetto di studio è utopico nella maggior parte dei casi. Ne consegue che quando si parla di corpora rappresentativi, si intende più un modello a cui tendere che un obiettivo che è possibile raggiungere in via definitiva.

Biber (1993) suggerisce una serie di indicazioni per ottenere corpora quanto più rappresentativi possibile. Il primo passo è definire la popolazione: bisogna decidere quali tipi di testo fanno parte della popolazione e quali non ne fanno parte. In secondo luogo, la popolazione va classificata e divisa in categorie. Tali categorie devono essere definite con chiarezza, in modo che, dato un testo, sia possibile assegnarlo alla categoria corrispondente. Inoltre, è necessaria una definizione operativa della popolazione per costruire un elenco degli individui su cui operare il campionamento. Infine, Biber (1993) sostiene che il campionamento stratificato proporzionale non sia adeguato alla costruzione di corpora. Questo perché un corpus non serve per calcolare statistiche descrittive della popolazione ma per analizzare la variabilità dei fenomeni linguistici e tale variabilità non viene catturata a pieno dal campionamento stratificato proporzionale.

I corpora più diffusi sono i corpora generali di riferimento. Il loro scopo è quello di rappresentare una lingua naturale nella sua varietà scritta o parlata. Tuttavia, questi corpora non documentano adeguatamente i linguaggi e i domini specialistici. Perciò c'è bisogno di corpora specialistici appositamente sviluppati che contengano un numero di testi e di tipi di testi sufficiente a rappresentare un dominio di conoscenza. (Sinclair, 1991).

Un corpus specialistico, infatti, contrariamente ad un corpus generale di riferimento, è composto principalmente da testi specialistici ed è realizzato appositamente per studiare le caratteristiche di un linguaggio speciale o la terminologia di uno specifico dominio di conoscenza.

3.2.1 Corpora in terminologia

Per la terminologia i corpora specialistici rappresentano una preziosa fonte di dati e informazioni. Questi corpora sono sempre stati fondamentali per la terminografia. Inoltre, oggi anche i terminologi fanno crescente affidamento sull'evidenza testuale per formulare ipotesi riguardo al comportamento dei termini in diversi contesti d'uso. In terminologia, i corpora specialistici trovano numerose applicazioni. Sono utilizzati per l'estrazione di termini attraverso gli algoritmi di EAT, forniscono statistiche sull'uso dei termini e sulla variazione terminologica e servono per individuare quei contesti che contengono definizioni o relazioni semantiche.

Sebbene sia difficile quantificare il contributo che i corpora e le applicazioni informatiche apportano al lavoro terminografico, la loro crescente diffusione suggerisce che essi offrano un grande supporto ai terminologi, permettendo loro di raccogliere maggiori informazioni sui termini e di classificarli in base a diversi parametri (L'Homme, 2006). Anche Cabré Castellví (2003, p. 189) sostiene che i corpora siano strumenti indispensabili. In particolare, afferma che:

> Se poniamo la terminologia nella cornice della comunicazione specialistica e riconosciamo che questa comunicazione avviene in diverse modalità, l'osservazione delle unità *[terminologiche]* può avvenire solo nei discorsi prodotti in queste situazioni. Il discorso scritto o orale degli specialisti indirizzato direttamente, o attraverso una qualche forma di mediazione, a specifici gruppi di destinatari costituisce il materiale nel quale possiamo osservare le unità terminologiche. Un corpus così diversificato costituirebbe un campione rappresentativo del discorso specialistico in tutte le sue forme.

Per essere una risorsa efficace, anche i corpora specialistici, come i corpora generali di riferimento devono essere costruiti seguendo determinati criteri che ne garantiscano il più possibile la qualità e la rappresentatività. Tuttavia, costruire un corpus è un'operazione complessa che richiede tempo e risorse e non è scevra da difficoltà pratiche, rendendo inevitabile la ricerca di un compromesso tra qualità e fattibilità (Bowker & Pearson, 2002).

In questa trattazione, definiamo due categorie di corpus specialistico: il corpus specialistico primario e il corpus specialistico secondario.

Corpus specialistico primario. Un corpus specialistico primario è un corpus composto da testi specialistici che non contiene un'annotazione manuale dei termini. In genere, un corpus specialistico primario è di grandi dimensioni e presenta un'annotazione automatica ottenuta attraverso *part-of-speech tagger*, lemmatizzatori e *parser* delle dipendenze. I corpora specialistici primari forniscono al terminologo una fonte di dati empirici per lo sviluppo di banche dati terminologiche e per lo studio dei linguaggi speciali.

Corpus specialistico secondario. Un corpus specialistico secondario è un corpus composto da testi specialistici in cui i termini sono stati annotati manualmente, in genere da due o più annotatori. Spesso, la finalità dei corpora specialistici secondari è quella di rappresentare un *gold-standard* per l'addestramento e la valutazione degli algoritmi di EAT. Sebbene siano ricchi di informazioni, di contro questi corpora tendono ad essere di dimensioni più contenute rispetto ai corpora specialistici primari.

La Tabella 3.1 offre una panoramica sui corpora specialistici, confrontando diversi corpora presentati in letteratura con i corpora sviluppati in questo lavoro: ItaIst-DdAC_GRU e ItaIst-TermRifiuti (evidenziati in grassetto).

3.2.2 Corpora specialistici primari

Una raccolta di corpora specialistici degna di nota è il progetto ParlaMint (Erjavec et al., 2024), una delle iniziative di punta di CLARIN.[1] ParlaMint è una raccolta di 29 corpora di dibattiti parlamentari provenienti da altrettanti paesi europei e regioni autonome, tra cui l'Italia. Questi corpora coprono un periodo che va dal 2015 al 2022 e contengono, in totale, oltre 1 miliardo di parole. Inoltre, la codifica dei dati è omogenea e include metadati dettagliati su circa 24.000 oratori, con annotazioni linguistiche che comprendono dipendenze sintattiche ed entità nominate.

Interessante è anche il corpus ACL ARC (Association for Computational Linguistics - Anthology Reference Corpus). ACL ARC è costituito da articoli scientifici in lingua inglese provenienti dall'ACL Anthology, un archivio digitale gestito dall'ACL che raccoglie la maggior parte della letteratura scientifica prodotta nell'ambito della linguistica computazionale e del TAL. Il corpus è quindi focalizzato sul dominio di conoscenza della linguistica computazionale e sui campi correlati. ACL ARC contiene 10.921 articoli e relativi metadati (autori, titolo, sede e anno di pubblicazione), con un totale di 152.546 riferimenti bibliografici. Il corpus ACL ARC nasce con lo scopo di facilitare lo sviluppo di algoritmi per l'estrazione automatica di informazioni bibliografiche da documenti accademici.

Come già accennato i corpora specialistici primari sono spesso utilizzati per sviluppare glossari, tesauri e basi di dati terminologiche. È il caso dell'EcoLexicon English Corpus (EEC) (Leoń-Araúz et al., 2018), un corpus specialistico di 23,1 milioni di parole dedicato a testi ambientali contemporanei. Questo corpus, insieme alla sua versione in spagnolo (EcoLexicon Spanish Corpus), è stato creato dal gruppo di ricerca LexiCon per supportare lo sviluppo della base di conoscenza EcoLexicon (Faber et al., 2016), Il corpus, inizialmente concepito come strumento interno per l'estrazione di conoscenze durante la costruzione di EcoLexicon è stato successivamente reso disponibile al pubblico per permettere a terminologi, traduttori ed esperti di sfruttarlo per diversi scopi. L'EEC include testi relativi a vari sottodomini dell'ambiente: biologia, meteorologia, ecologia, ingegneria ambientale, diritto ambientale, e altri. I testi sono categorizzati secondo vari parametri, come sottodominio, tipo di lettori (esperti, semi-esperti, pubblico generale), variante geografica (inglese americano, inglese britannico, ed *euro-english*), genere (articoli, libri, siti web, ecc.), editore (accademico, aziendale, governativo), anno di pubblicazione (dal 1973 al 2016) e paese di pubblicazione.

Per quanto riguarda i corpora specialistici primari in lingua italiana, segnaliamo i corpora Univers-ITA, PAC & RAC, CultHer e PILLAR. Il corpus Univers-ITA (Grandi et al., 2023; Restivo, 2022) è composto da 2160 testi argomentativi redatti da studenti universitari che frequentano il secondo anno accademico di corsi di laurea triennale o a ciclo unico. Nello specifico, agli studenti è stato chiesto di redigere un breve testo (circa 500 parole) che esprimesse la loro opinione sulla didattica a distanza. Univers-ITA nasce con lo scopo di indagare le competenze di scrittura degli studenti universitari. Per questo motivo, ogni testo è corredato di un ricco set di metadati relativi all'autore, che sono stati raccolti attraverso un questionario socio-biografico di circa 50 domande.

[1] CLARIN è un'infrastruttura europea di ricerca che offre dati, strumenti e servizi per supportare la ricerca basata sulle risorse linguistiche. Cfr. https://www.clarin.eu/ (Ultima rilevazione 15 maggio 2025).

I corpora PAC (*Pro Ana/Anorexia Corpus*) e RAC (*Recovery from Ana/Anorexia Corpus*) (Donati et al., 2023) raccolgono le trascrizioni di 1250 video TikTok relativi a disturbi alimentari e all'anoressia. In particolare, il corpus PAC è composto da 250 trascrizioni di video in lingua inglese a favore dell'anoressia.[2] Al contrario, le 1000 trascrizioni in lingua italiana, contenute in RAC, riguardano il processo di guarigione dall'anoressia.

Il corpus CultHer (Cimmino et al., 2022) contiene un'ampia raccolta di testi dedicati al settore turistico, focalizzati sul concetto di turismo emozionale. Questi testi, oltre a descrivere i luoghi turistici e gli eventi legati al turismo, includono anche le opinioni degli utenti raccolte attraverso recensioni, racconti di viaggio e spot pubblicitari. Per quanto riguarda i tipi di testo, CultHer comprende testi descrittivi, che dettagliano i luoghi e le esperienze turistiche, testi informativi che forniscono notizie e aggiornamenti sul settore, e testi argomentativi che discutono e promuovono il concetto di turismo emozionale.

Il corpus PILLAR (Parallel ItaLian engLish ARchaeological corpus) (Speranza et al., 2023) è un corpus parallelo italiano-inglese focalizzato sul dominio dell'archeologia che nasce per l'EAT. Nello specifico, il corpus è composto da 20 testi (circa 200.000 token per ciascuna lingua), suddivisi in brochure (4), guide di musei e siti archeologici (12) e siti web (4). I testi sono stati selezionati per garantire una qualità elevata e una corrispondenza completa tra la versione italiana e quella inglese, pertanto provengono da fonti autorevoli e coprono un periodo temporale che va dal 2006 al 2020.

3.2.3 Corpora specialistici secondari

Molti corpora specialistici secondari nascono nell'ambito dell'EAT, dove sono utilizzati per valutare l'accuratezza degli algoritmi e addestrare modelli supervisionati. Tra i principali corpora di questo tipo annoveriamo GENIA (Kim et al., 2003) ACL RD-TEC (Qasemizadeh & Schumann, 2016) e ACTER (Rigouts Terryn et al., 2020).

GENIA (Kim et al., 2003) è un corpus composto da 2000 abstract in lingua inglese, estratti dal database MEDLINE. Il dominio è quello della biologia. Nello specifico, il corpus riguarda i fattori di trascrizione nelle cellule del sangue umano. Nel corpus GENIA, ogni termine è un'istanza di una classe definita nell'omonima ontologia. Ad esempio, il termine *tumore fibroblastico* è un'istanza della classe *tessuto*.

Il corpus ACL RD-TEC (Qasemizadeh & Schumann, 2016) è composto da 300 abstract, sempre in lingua inglese, tratti dall'ACL Anthology Reference Corpus. ACL RD-TEC è stato sviluppato con l'obiettivo di creare un corpus per l'EAT per il quale i linguisti computazionali siano essi stessi esperti di settore. In questo modo, possono sviluppare e testare i loro algoritmi senza aver bisogno dell'aiuto di esperti esterni. In questo corpus, i termini sono classificati sulla base di linee guida sviluppate durante le prime fasi di annotazione.

ACTER (Rigouts Terryn et al., 2020) è un corpus terminologico multilingue e multidominio. Esso copre quattro diversi domini di conoscenza (corruzione, equitazione, insufficienza cardiaca, ed energia eolica) e contiene testi in tre diverse lingue: inglese francese e olandese. ACTER è stato appositamente costruito per testare gli algoritmi di EAT in un contesto multidominio e multilingue. Esso mantiene un'annotazione coerente tra lingue e domini, permettendo così di comparare i risultati. In ACTER, i termini sono classificati secondo uno schema composto da quattro etichette: termini

[2] PAC è in lingua inglese poiché gli autori dello studio non hanno trovato materiale pro-anoressia in italiano.

specialistici, termini comuni, termini *out-of-domain* e non-termini. Inoltre, esso distingue i termini dalle entità nominate.

Per quanto riguarda le risorse in lingua italiana, segnaliamo BitterCorpus (Arcan et al., 2014) e MAGMATic (Scansani et al., 2019), due corpora paralleli italiano-inglese in cui le unità terminologiche sono annotate manualmente. BitterCorpus contiene testi provenienti da corpora preesistenti sul dominio dell'*information technology* mentre MAGMATic è composto da descrizioni di corsi universitari riguardanti materie scientifiche. Entrambi i corpora nascono nell'ambito della valutazione della traduzione automatica dei testi specialistici.

Tipo	**Nome**	**Dominio**	**Lingua**	**Testi**	**Token**
primario	**ItaIst-DdAC_GRU**	gestione dei rifiuti	italiano	306 documenti	1.351.921
	ParlaMint	dibattiti parlamentari	29 lingue	-	1.231.036.093
	Ecolexicon English Corpus	ambiente	inglese	2.214 documenti	28.616.037
	Univers-ITA	opinioni sulla didattica a distanza	italiano	2.137 testi	893.051
	PAC e RAC	disturbi dell'alimentazione e anoressia	italiano; inglese	1.250 trascrizioni	129.430
	CultHer	turismo emozionale	italiano	737 testi	270.188
	PILLAR	archeologia	inglese; italiano	20 testi	220.000
secondario	**ItaIst-TermRifiuti**	gestione dei rifiuti urbani	italiano	campioni da 69 documenti	41.333
	GENIA	biologia*	inglese	2.000 abstract	436.967
	ACL RD-TEC	TAL	inglese	300 abstract	33.216
	ACTER	corruzione; equitazione; insufficienza cardiaca; energia eolica	inglese; francese; olandese	1.194 documenti	596.058
	BitterCorpus	*information technology*	inglese; italiano	55 documenti + 100 frasi	-
	MAGMATic	corsi universitari	inglese; italiano	2,157 frasi	36.162

Tabella 3.1: Corpora specialistici

3.2.4 Formati di rappresentazione

Per poter essere studiati con maggiore efficacia, i corpora dovrebbero essere compatibili sia con altri corpora che con risorse lessicografiche e software di TAL. Per questo motivo è fondamentale adottare dei formati di rappresentazione che siano standard e interoperabili. I formati standard più diffusi per rappresentare i corpora sono tre: CoNLL-U[3] (*Conference on Natural Language Learning - Universal*), TEI/XML (*Text Encoding Initiative / eXtensible Markup Language*) (TEI, 2025) e NIF (*NLP Interchange Format*) (Hellmann et al., 2013).

[3]https://universaldependencies.org/format.html, (Ultima rilevazione: 27/02/2025)

CoNLL-U. CoNLL-U è un formato testuale di tipo *Tab Separated Value* (TSV) che utilizza il carattere di tabulazione per separare le colonne, un singolo carattere di invio per separare le righe e due caratteri di invio per separare le frasi. In un file CoNLL-U. ogni riga rappresenta un token e le colonne ne descrivono le proprietà: forma, lemma, parte del discorso, dipendenze sintattiche, e altro. Il formato CoNLL-U nasce dal progetto *Universal Dependencies*, che mira a migliorare la compatibilità tra risorse linguistiche in diverse lingue e standardizzare le categorie e le linee guida di annotazione, per facilitare l'annotazione di fenomeni cross-linguistici. CoNLL-U è probabilmente il formato più utilizzato nei compiti di TAL proprio perché, nonostante la sua semplicità, è in grado di rappresentare le principali proprietà dei *token* in diverse lingue.

Il maggiore limite di questo formato risiede nel suo stretto legame con la divisione in *token* del testo. In un file CoNLL-U la divisione in *token* è obbligatoria e non è possibile dividere i *token* secondo criteri diversi (Cimiano et al., 2020).

TEI/XML. Un altro standard molto diffuso è il formato TEI/XML. È basato sul formato XML e segue le linee guida stabilite dal consorzio *Text Encoding Initiative* (TEI). TEI/XML permette di rappresentare la struttura logica di un testo (titoli, paragrafi, capitoli, note a piè di pagina, ecc.) e di aggiungere metadati dettagliati per descrivere elementi come autori, date, fonti, e relazioni semantiche. Inoltre, TEI/XML può essere adattato a diversi tipi di testi, come manoscritti, edizioni critiche, archivi storici, documenti giuridici e altro. Il progetto TEI nasce in ambito umanistico, con lo scopo di rappresentare in formato elettronico documenti storici e letterari. Tuttavia, grazie alla sua flessibilità, il formato TEI/XML è adatto a rappresentare qualsiasi tipo di documento ed ha una discreta diffusione anche nell'ambito del TAL.

NIF. *NLP Interchange Format* (NIF) (Hellmann et al., 2013) è un'ontologia OWL che permette di rappresentare corpora testuali in formato RDF (*Resource Description Framework*), nel contesto del web semantico. Il punto di forza di NIF sono gli schemi *Uniform Resource Identifier* (URI) tramite i quali è possibile assegnare un URI a una porzione di testo. Ad esempio nello schema URI *offset-based*, una porzione di testo è rappresentata come nel modello seguente.

```
{URI del documento}#offset_{indice di inizio}_{indice di fine}
es. https://example.com/doc1#offset_5_15
```

Grazie agli URI è possibile assegnare vari livelli di annotazione a porzioni di testo di diverse dimensioni. L'annotazione della struttura logica del testo è rappresentata direttamente tramite NIF mentre le annotazioni linguistiche sono rappresentate tramite le (*Ontologies of Linguistic Annotation*) (OLiA).[4] Il limite di NIF è che esso non permette la rappresentazione di porzioni di testo discontinue ma è possibile sopperire a questa limitazione utilizzando l'ontologia *Portable Linguistic Annotation with OWL* (POWLA) (Chiarcos, 2012b) come suggerito in (Cimiano et al., 2020). Sebene NIF non sia ufficialmente uno standard, esso rappresenta uno standard *de-facto*, essendo uno dei formati più utilizzati per riferirsi a entità di linguaggio naturale nel contesto del web semantico (Cimiano et al., 2020).

Il formato NIF presenta diversi vantaggi rispetto ai formati CoNLL-U e TEI/XML. Innanzitutto non è legato alla tokenizzazione come il formato CoNLL-U ma preserva l'integrità del testo originale

[4] https://acoli-repo.github.io/olia/ (Ultima rilevazione: 28/02/2025)

che, di fatto, è mantenuto separato dall'annotazione. Inoltre, l'uso di URI garantisce l'interoperabilità delle risorse e permette di collegare i corpora con qualsiasi altro tipo di dato in formato RDF, sia esso linguistico o meno. Per questo motivo NIF è stato scelto per rappresentare i corpora specialistici ItaIst-DdAC_GRU e ItaIst-TermRifiuti e per collegarli alla risorsa terminologica ItaIst-WasteLexicon.

3.3 Estrazione automatica di termini

L'Estrazione Automatica di Termini (EAT) è un compito di TAL che consiste nell'individuare automaticamente i termini presenti in un corpus specialistico. Gli algoritmi di EAT tradizionali hanno un'architettura simile che prevede tre fasi: l' estrazione dei candidati, il calcolo delle *feature* e la stima della probabilità (Astrakhantsev et al., 2015).

Più di recente, con l'evolversi del TAL, sono emersi nuovi approcci, basati sul *deep learning* e sul *prompting*, per i quali lo schema in tre fasi non trova applicazione. Infatti questi approcci non necessitano del calcolo di metriche esplicite ma si basano piuttosto su rappresentazioni distribuite di parole e frasi interne ai *Large Language Model* (LLM).

3.3.1 Approcci tradizionali

Estrazione dei candidati. La fase di estrazione dei candidati consiste nell'estrarre da un corpus specialistico i candidati termini, cioé parole e sequenze di parole ben formate, che potrebbero essere termini. In questa fase viene prodotta una lista di candidati che viene poi rielaborata nelle fasi successive.

La tecnica più comune per estrarre i candidati prevede l'impiego di pattern morfosintattici. Nello specifico, vengono estratte quelle parole e sequenze di parole che hanno determinate caratteristiche morfologiche e sintattiche. I pattern più utilizzati sono quelli che estraggono semplici sintagmi nominali, come ad esempio `<Nome>`, `<Nome + Aggettivo>` e `<Nome + Preposizione + Nome>`. In genere, questi pattern sono predefiniti (Justeson & Katz, 1995; Meyers et al., 2018) ma possono anche essere inferiti a partire da una lista di termini già noti (Pais & Ion, 2020). Un altra tecnica di estrazione dei candidati è il *chunking* che consiste nel separare le frasi in gruppi di parole considerando la punteggiatura e le *stop-words* come delimitatori. (Rose et al., 2010). Inoltre la tecnica *Token Slot Recognition* filtra ulteriormente i candidati, selezionando solo quelli composti da parole che occorrono in termini già noti (Oliver & Vàzquez, 2020).

Calcolo delle *feature*. Il calcolo delle *feature* consiste nel calcolare, per ogni candidato termine, alcune metriche che catturano i tratti caratteristici dei termini e li distinguono dal resto del lessico.

La maggior parte delle metriche è di tipo statistico e sintetizza la distribuzione dei candidati termini in un corpus specialistico. La metrica più semplice è la *Term Frequency* ed è uguale al numero di occorrenze del candidato termine nel corpus specialistico. Altre metriche esaminano anche la distribuzione dei candidati termini nei documenti del corpus (Church & Gale, 1999; Evans & Lefferts, 1995; Navigli & Velardi, 2004). Esiste, poi, un altro tipo di metriche, anch'esse molto diffuse, che comparano la frequenza dei candidati termini nel corpus specialistico con la loro frequenza in un corpus generale di riferimento. (Ahmad et al., 1999; Kilgarriff, 2009; Meijer et al., 2014; Park et al., 2002)

Infine, esistono delle metriche semantiche, che utilizzano i *word embedding* per misurare la similarità semantica tra i candidati termini. Una di queste, la *Key Concept Relatedness*[5] (Astrakhantsev, 2018), misura la similarità tra i candidati e dei termini già noti del dominio di riferimento. Inoltre, Pais e Ion (2020) hanno utilizzato i *word embeddings* e un algoritmo di *clustering*, DBSCAN, per raggruppare i candidati termini dello stesso dominio. Questa tecnica ha ottenuto prestazioni migliori degli approcci statistici. Alcune delle metriche menzionate, sono implementate nei software JATE 2.0 (Zhang et al., 2016) e ATR4S (Astrakhantsev, 2018).

Stima della probabilità. Infine, nella fase di stima della probabilità, viene calcolata per ciascun candidato la sua probabilità di essere un termine. Questo passaggio è superfluo quando è coinvolta una sola *feature*, poiché le probabilità dipendono unicamente dal valore di tale *feature*.

Viceversa, per tenere conto di più *feature*, esistono diverse strategie di combinazione, la più semplice delle quali è la *media ponderata* (Ahmad et al., 1999). Un'altra strategia è l'algoritmo di votazione, che considera solo il rango dei candidati invece dei punteggi effettivi (Fedorenko et al., 2014; Pais & Ion, 2020). Inoltre, quando è disponibile un dataset *gold-standard* con etichette di riferimento, è possibile utilizzare algoritmi di *machine learning* e addestrare un classificatore supervisionato (Patry & Langlais, 2005).

3.3.2 Approcci basati su *deep learning* e *prompting*

I modelli di *deep learning* sono reti neurali composte da numerosi strati di neuroni artificiali che possono essere addestrate a svolgere un determinato compito sulla base di un *dataset* di esempi, solitamente di grandi dimensioni. Negli ultimi anni questi modelli hanno dimostrato performance eccellenti in numerosi i compiti di TAL, compresa l'EAT. Uno dei modelli di *deep learning* più diffuso è sicuramente BERT (*Bidirectional Encoder Representations from Transformers*) (Devlin et al., 2019), un modello preaddestrato basato su reti neurali *transformer* che può essere specializzato, attraverso il *fine-tuning* per svolgere uno specifico compito di TAL anche con un *dataset* relativamente piccolo.

Infine, l'ultima frontiera del TAL sono i *Large Language Model instruction tuned*, modelli generativi che sono poi stati istruiti, tramite il *fine-tuning* ad eseguire istruzioni di vario genere, dette *prompt*. Per utilizzare questi modelli in un compito TAL, è sufficiente formulare il compito in un *prompt* nel quale viene inserito il testo da trattare.

Per effettuare l'EAT, Kucza et al. (2018) propongono un modello LSTM che esegue il *sequence labelling* utilizzando uno schema BILOU (*Beginning*, *Inside*, *Last*, *Outside*, *Unit*). Invece, Hazem et al. (2020), tramite il *fine-tuning* del modello BERT (Devlin et al., 2019), hanno ottenuto risultati notevoli nell'EAT, superando le performance dei modelli precedenti. Manjunath e McCrae (2021), invece, hanno utilizzato un modello BERT per produrre i *word-embedding* contestuali da dare in input a un modello di *deep learning* appositamente progettato per l'EAT. Lang et al. (2021) hanno sviluppato e testato due architetture basate sul modello multilingue XLM-RoBERTa (Conneau et al., 2020). La prima architettura struttura l'EAT come un problema di classificazione mentre la seconda la struttura come un problema di *sequence labeling*. Quest'ultima architettura ha ottenuto risultati migliori.

[5] La *Key Concept Relatedness* è descritta più nel dettaglio nel Capitolo 5

Sebbene siano, fino ad ora, i metodi più efficaci per l'EAT, i metodi di *deep learning* presentano un rilevante svantaggio: necessitano di essere addestrati su dataset *gold-standard* che, purtroppo, non sempre sono disponibili per la lingua e il dominio trattati.

Per affrontare questo problema, Banerjee et al. (2024) hanno comparato GPT-3.5-Turbo (Brown et al., 2020) a XLM-RoBERTa (Conneau et al., 2020) simulando uno scenario in cui gli esempi su cui addestrare il modello sono limitati (*few-shot learning*). I risultati hanno mostrato come GPT-3.5-Turbo ottenga buoni risultati anche con pochi esempi, mentre XLM-RoBERTa ha bisogno di numerosi esempi per eguagliare le performance di GPT-3.5-Turbo. Anche Tran et al. (2024) hanno testato diversi LLM giungendo alla conclusione che, nei contesti in cui i dati sono scarsi, le tecniche di *prompting* funzionano meglio di quelle di *deep learning*.

3.3.3 Tecniche per isolare uno strato terminologico

La maggior parte degli approcci all'EAT, compresi quelli illustrati in precedenza, estrae indiscriminatamente tutti i termini presenti in un corpus specialistico, indipendentemente dal dominio a cui appartengono. Invece, l'obiettivo della *Domain Concept Relatedness*, la tecnica di EAT proposta nel Capitolo 5, è quello di isolare la terminologia di un determinato strato terminologico. Attualmente, esistono almeno altri due metodi che hanno lo stesso obiettivo. Essi sono l'approccio contrastivo, proposto da Basili et al. (2001) e Bonin et al. (2010) e l'approccio proposto da Drouin et al. (2018b).

L'ipotesi alla base dell'approccio contrastivo è che la distribuzione di un candidato nel corpus specialistico esaminato (detto *corpus di riferimento*) non sia da sola sufficiente per determinarne la sua appartenza al dominio considerato. È necessario anche esaminarne la distribuzione in corpora specialistici relativi ad altri domini (detti *corpora contrastivi*). Per calcolare il peso contrastivo dei candidati, Basili et al. (2001) propongono due metriche: una per i termini monorematici e una per i polirematici. Nello specifico, Il peso contrastivo dei termini monorematici è uguale al prodotto della loro frequenza nel corpus di riferimento e l'*Inverse Word Frequency* in tutti i corpora considerati. Essendo i termini polirematici meno frequenti, la loro probabilità è più difficile da stimare. Per questo motivo, il peso contrastivo dei termini polirematici è ottenuto moltiplicando il peso contrastivo della loro testa sintattica per la frequenza dell'intero termine nel corpus di riferimento. Invece, Bonin et al., 2010 utilizzano la stessa formula di calcolo sia per i termini monorematici che per quelli polirematici, grazie ad una funzione addatta alla stima di eventi a bassa frequenza.

Drouin et al. (2018b) tentano di isolare lo strato terminologico di un dominio molto ampio, ovvero il dominio dell'ambiente, per catturare il *lessico ambientale generale* e separarlo dalla terminologia di temi più specifici come quelli delle energie rinnovabili, della gestione dei rifiuti e delle specie a rischio. In particolare, testano due metriche: la *specificità* e l'*Inverse Document Frequency* (IDF). La specificità compara la distribuzione dei termini nel corpus specialistico con la loro frequenza in un corpus generale di riferimento mentre l'IDF misura la distribuzione dei termini nei singoli documenti. Drouin et al. (2018b) si aspettano che i termini appartenenti al *lessico ambientale generale* ottengano un'alta specificità, perché sono caratteristici dei discorsi specialistici riguardanti l'ambiente ma una bassa IDF, perché sono equidistribuiti tra i documenti che trattano temi specifici.

3.4 Gestione e pubblicazione delle risorse terminologiche

Le risorse terminologiche sono il risultato del lavoro terminografico: raccolte di dati terminologici organizzate secondo criteri ben precisi che spesso assolvono la funzione di opere di consultazione. La norma ISO 1087:2019 definisce il concetto di risorsa terminologica in questo modo:

> **Risorsa terminologica:** Raccolta di *voci terminologiche*.
>
> **Voce terminologica:** Raccolta di *dati terminologici* relativi a un solo concetto.
>
> **Dati terminologici:** Dati relativi a concetti e alle loro designazioni. I dati terminologici comuni includono designazioni, definizioni, contesti, note alle voci, etichette grammaticali, etichette relative alla materia, identificativi di lingua, identificativi di paese e identificativi di fonte.

La stessa norma individua tre tipi principali di risorse terminologiche:

- I vocabolari, che contengono i termini con le relative definizioni.
- I glossari, che contengono i termini e i loro equivalenti di traduzione.
- Le nomenclature, che sono strutturate in modo sistematico secondo regole di denominazione prestabilite.

3.4.1 Formati di rappresentazione

Oltre al tradizionale formato cartaceo, le risorse terminologiche vengono sempre più spesso realizzate anche in formato elettronico. In questo modo, oltre ad essere impiegate come opere di consultazione, possono essere integrate in diverse applicazioni che vanno dalla traduzione automatica, alla semplificazione di testi.

A tal fine, il formato di rappresentazione deve garantire che le risorse terminologiche siano compatibili tra di loro e con risorse di altro tipo, per cui è fondamentale adottare formati di rappresentazione che siano standard e interoperabili. Attualmente, i formati più diffusi tra i terminografi sono tre: *TermBase eXchange format* (TBX) ISO 30042:2019, *Simple Knowledge Organization System* (SKOS) (W3C, 2009) e *OntoLex-Lemon* (Cimiano et al., 2016; McCrae et al., 2017).

TBX. TBX è uno standard internazionale basato su XML che nasce per favorire lo scambio di risorse terminologiche tra diversi software di gestione terminologica e traduzione assistita. In TBX, una risorsa terminologica è una collezione di dati terminologici che contiene una o più entrate (`conceptEntry`). Ogni entrata si riferisce ad un solo concetto e include una o più sezioni di lingua (`langSec`) tramite le quali il concetto è espresso. Poi, per ogni sezione di lingua ci sono una o più sezioni terminologiche (`termSec`) che contengono i termini e le informazioni ad essi correlate, come ad esempio la parte del discorso.

SKOS. SKOS è un'ontologia OWL che nasce per rappresentare, nel contesto del web semantico, diversi modelli di organizzazione della conoscenza, quali tesauri, schemi di classificazione, tassonomie e altri tipi di vocabolari controllati. L'elemento centrale del modello SKOS è il concetto, istanza della classe `Concept`. Ad un concetto può essere associata un'etichetta preferita (`prefLabel`) per ogni lingua e un numero qualsiasi di etichette alternative (`altLabel`). Inoltre, ad ogni concetto possono

essere associati altri dati come note (`note`), esempi (`example`) e definizioni (`definition`). Infine, SKOS permette di definire le relazioni tra concetti. Le relazioni gerarchiche sono rappresentata tramite le proprietà `broader` e `narrower` mentre le relazioni non gerarchiche sono rappresentate tramite la proprietà `related`. Esiste poi un'estensione di SKOS, SKOS-XL (Miles & Bechhofer, 2009) che introduce, tramite etichette reificate, la possibilità di aggiungere metadati alle etichette e di metterle in relazione.

Ontolex-Lemon. Anche OntoLex-Lemon è un'ontologia OWL che permette di rappresentare informazioni linguistiche nel formato del web semantico. Lo scopo originale di OntoLex-Lemon non è quello di rappresentare risorse terminologiche, bensì quello di formalizzare i dati linguistici relativi a ontologie e vocabolari controllati. Nonostante ciò, OntoLex-Lemon ha trovato tra i terminologi una diffusione molto ampia (Fiorelli et al., 2018; Martín-Chozas & Declerck, 2022; Piccini et al., 2023) e ne è stata suggerita un'estensione specifica per la terminologia.[6] In OntoLex-Lemon, i dati terminologici sono rappresentati attraverso quattro elementi principali: le entrate, le forme, i concetti e i sensi. Le entrate sono istanze della classe `LexicalEntry` e rappresentano unità linguistiche, in questo caso termini. Le forme sono istanze della classe `Form` e rappresentano le forme flesse di un'entrata. I concetti sono istanze della classe `LexicalConcept` e rappresentano unità di conoscenza. Infine, i sensi, che sono istanze della classe `LexicalSense`, rappresentano il significato delle entrate e le collegano ai concetti. Lo schema in Figura 3.1 rappresenta la formalizzazione in OntoLex-Lemon delle due entrate terminologiche *centro di raccolta* e *isola ecologica*.

Il formato OntoLex-Lemon presenta diversi vantaggi rispetto al formato TBX (Piccini 2023). Innanzitutto garantisce l'interoperabilità semantica, oltre che quella sintattica. Inoltre, l'uso di URI in OntoLex-Lemon permette di connettere le risorse terminologiche con qualsiasi altro tipo di dati strutturati in formato RDF, siano essi linguistici o di altro tipo. Infine, OntoLex-Lemon, a differenza di TBX, permette di formulare regole di inferenza che arricchiscono i dati di ulteriori informazioni. Ad esempio si può inferire che se due termini denotano il medesimo concetto nella stessa lingua essi sono sinonimi. Per questo motivo OntoLex-Lemon è stato scelto come formato di rappresentazione per sviluppare la risorsa terminologica ItaIst-TermWaste e collegarla al corpus ItaIst-TermRifiuti.

3.4.1.1 Editor di risorse terminologiche

Data la complessità dei formati TBX e OntoLex-Lemon, è poco pratico sviluppare e tenere aggiornate risorse terminologiche intervenendo direttamente sui file sorgente. Inoltre, i terminografi con una formazione umanistica hanno generalmente scarsa familiarità con questi formati.

Per questo motivo sono disponibili diversi editor che permettono ai terminografi di lavorare alle risorse terminologiche senza doversi preoccupare dei complessi formalismi sottostanti. In questo paragrafo analizzeremo brevemente i principali editor di risorse terminologiche *open-source*: FAIRTerm (Vezzani, 2021), che tratta il formato TBX, VocBench (Stellato et al., 2020, 2015), LexO (Bellandi, 2021) e Maia (Giovannetti et al., 2024) che trattano il formato OntoLex-Lemon.

FAIRTerm. FAIRTerm (Vezzani, 2021) è un editor di risorse terminologiche progettato per supportare la didattica e la professionalizzazione nel campo della traduzione specialistica. Una scheda terminologica multilingue di FAIRTerm contiene 40 campi compilabili, divisi in quattro pannelli:

[6] https://www.w3.org/community/ontolex/wiki/Terminology (Ultima rilevazione: 28/01/2024)

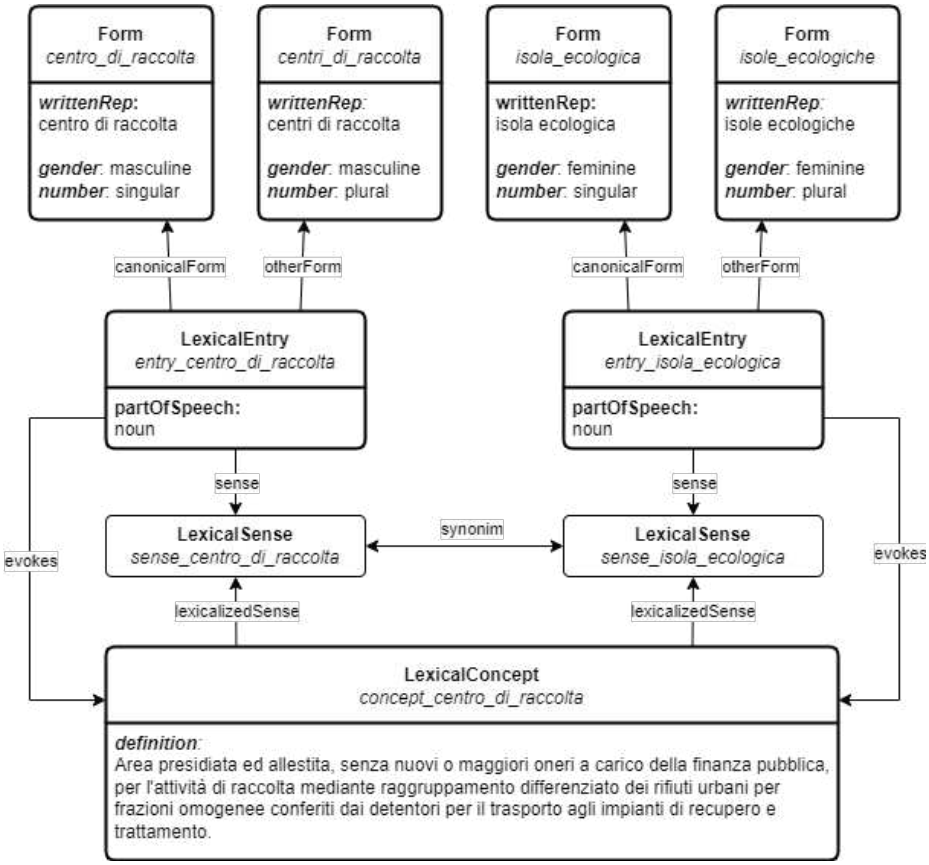


Figura 3.1: Entrate terminologiche rappresentate in OntoLex-Lemon

caratteristiche formali, *semantica*, *variazione* e *uso*. Il pannello *caratteristiche formali* contiene prevalentemente informazioni morfo-grammaticali, tra cui parte del discorso, genere e numero. Il pannello *semantica* contiene informazioni sul significato del termine, tra cui definizione, fonte e sinonimi. Il pannello *variazione*, invece, contiene informazioni sulle varianti terminologiche, tra cui nome comune, nome scientifico, acronimi e abbreviazioni. Infine, il pannello *uso* contiene informazioni sulle situazioni comunicative in cui il termine è impiegato: dominio, sottodominio, contesti, ecc. Nel pieno rispetto del paradigma FAIR (Wilkinson et al., 2016), le schede terminologiche di FAIRTerm sono esportabili in formati compatibili con i sistemi di traduzione assistita, come il formato TBX.

VocBench. VocBench (Stellato et al., 2020, 2015) è una versione migliorata della piattaforma Agrovoc workbench, un editor appositamente sviluppato all'interno dell'Organizzazione delle Nazioni Unite per l'alimentazione e l'agricoltura (FAO) per lavorare sul tesauro Agrovoc (Subirats-Coll et al., 2022). Agrovoc workbench permetteva di lavorare solo ad Agrovoc, mentre VocBench permette di gestire qualsiasi tesauro SKOS. La versione attuale della piattaforma, VocBench 3 (Stellato et al., 2020), supporta il formato OWL e le ontologie SKOS, SKOS-XL e OntoLex-Lemon. Uno dei punti di forza di VocBench è il sistema di gestione dei ruoli che assegna specifici permessi agli utenti. In particolare, tra i ruoli standard ci sono quelli di:

- *Amministratore:* possiede tutti i permessi su tutti i progetti.
- *Direttore di progetto:* possiede tutti i permessi su un solo progetto.
- *Editor di ontologie:* può modificare solo le ontologie OWL.
- *Editor di tesauri:* può modificare solo i tesauri SKOS.
- *Terminografo/lessicografo:* può modificare solo i termini associati ai tesauri.
- *Visualizzatore:* può visualizzare tutto il progetto ma non può modificarlo.

Sebbene l'interfaccia di VocBench sia abbastanza intuitiva, la piattaforma è pensata per chi ha già una certa familiarità con i formati del web semantico e non è quindi adatta a tutti i tipi di utenti.

LexO. Sviluppato specificamente nell'ambito dei progetti umanistici, l'obiettivo di LexO (Bellandi, 2021) è quello di rendere OntoLex-Lemon fruibile agli utenti che non hanno familiarità con le tecnologie del web semantico. I dati di LexO, così come quelli di VocBench, possono essere esportati nei formati RDF/XML e Turtle, che sono standard per la rappresentazione di informazioni nel contesto del web semantico. Il vantaggio fondamentale di LexO risiede nella sua semplicità d'uso. Grazie a un'interfaccia intuitiva, LexO democratizza l'accesso al modello OntoLex-Lemon rendendolo accessibile anche a quegli utenti che desiderano sviluppare risorse interoperabili secondo il modello del web semantico ma sono sprovvisti di competenze tecniche avanzate.

Maia. Maia (Giovannetti et al., 2024) è uno strumento integrato che combina l'annotazione dei testi e la costruzione di risorse lessicali in un unico ambiente. Le funzionalità principali di Maia includono la compilazione di risorse lessicali e la costituzione di corpora con annotazioni multilivello. Inoltre, Maia permette di visualizzare le concordanze e collegare le annotazioni alle risorse lessicali. Per gestire le risorse lessicali e lessicografiche MAIA utilizza LexO-server, il *backend open-source*

di LexO-lite. Maia consente di esportare in RDF i lessici rappresentati tramite OntoLex-Lemon. Infine, è stato previsto di introdurre in Maia la possibilità di esportare in RDF anche le annotazioni, rappresentate tramite le ontologie Web Annotation e POWLA.

3.4.2 Principali banche dati terminologiche istituzionali

Una delle principali banche dati terminologiche istituzionali è sicuramente *InterActive Terminology for Europe* (IATE[7]), la banca dati terminologica multilingue dell'Unione Europea. IATE è attualmente una delle più ricche banche dati terminologiche esistenti: contiene circa 8 milioni di termini in 24 lingue ufficiali dell'Unione Europea e copre una vasta gamma di settori specialistici. IATE nasce dal consolidamento di diverse banche dati istituzionali preesistenti, tra cui Eurodicautom, TIS, Euterpe, Euroterms e CDCTERM.

In IATE, una scheda terminologica è composta da diverse sezioni: il termine, la lingua in cui è espresso, la definizione, esempi di contesto e la fonte da cui derivano le informazioni (documenti ufficiali o pubblicazioni specializzate). Inoltre, ogni voce è accompagnata da un indicatore di affidabilità, rappresentato da un sistema di valutazione a stelle. L'accesso pubblico e gratuito a IATE rappresenta una risorsa fondamentale non solo per le istituzioni dell'Unione Europea, ma anche per studiosi e ricercatori interessati alla terminologia specifica dell'ambito comunitario.

Un'altra base di dati fondamentale per la terminologia istituzionale dell'Unione Europea è Eurovoc,[8] un vocabolario controllato multilingue il cui scopo principale è quello di classificare i documenti presenti in EUR-Lex[9] in base al loro contenuto. Eurovoc contiene parole chiave in 23 lingue, organizzate in 21 settori e 127 sottosettori.

Invece, per quanto riguarda l'italiano istituzionale svizzero, la risorsa di riferimento è Termdat,[10] la banca dati terminologica dell'amministrazione federale svizzera che contiene i termini istituzionali di numerosi settori nelle quattro lingue nazionali svizzere (il tedesco, il francese, l'italiano e il romancio) e spesso anche in inglese. In Termdat, ogni scheda terminologica rappresenta un concetto e contiene informazioni circa i termini utilizzati per esprimerlo e i loro contesti d'uso. Le schede terminologiche Termdat sono redatte dagli uffici di terminologia di diverse istituzioni svizzere.

Anche la Provincia Autonoma di Bolzano ha sviluppato, in collaborazione con l'Istituto di Linguistica Applicata di Eurac Research, la propria banca dati terminologica, chiamata Bistro.[11] Bistro è finalizzata alla standardizzazione della terminologia legale ed amministrativa nelle tre lingue della provincia: tedesco, italiano e ladino. Inoltre, include i termini standardizzati dalla Commissione di Terminologia dell'Alto Adige, istituita nel 1994 per uniformare la terminologia giuridica e amministrativa in tedesco nella Provincia di Bolzano, eliminando varianti terminologiche incerte e migliorando la coerenza linguistica. In Bistro, le schede terminologiche sono organizzate secondo una struttura onomasiologica che permette di accedere alle informazioni linguistiche, catalografiche e giuridiche di ciascun termine.

[7] https://iate.europa.eu/home (Ultima rilevazione: 21/03/2025)
[8] https://eur-lex.europa.eu/browse/eurovoc.html?locale=it (Ultima rilevazione: 21/03/2025)
[9] EUR-Lex è il portale ufficiale dell'Unione Europea che fornisce accesso alla legislazione UE.
[10] https://www.termdat.bk.admin.ch/search (Ultima rilevazione: 21/03/2025)
[11] https://bistro.eurac.edu/ (Ultima rilevazione: 21/03/2025)

Capitolo 4

Corpora

Contenuti

4.1 Introduzione 39
4.2 ItaIst-DdAC_GRU 40
4.2.1 Raccolta e metadatazione dei testi 40
4.2.2 Estrazione del testo e preprocessing 42
4.2.3 Composizione del corpus 42
4.3 ItaIst-TermRifiuti 43
4.3.1 Accordo tra annotatori 45
4.4 Formati di rappresentazione e modalità di interrogazione 47

4.1 Introduzione

Lo studio della terminologia istituzionale della gestione dei rifiuti urbani non può prescindere da una solida base di dati empirici, ovvero un corpus di testi istituzionali che sia il più possibile rappresentativo dei discorsi specialistici sul tema della gestione dei rifiuti urbani. In particolare, tale corpus deve rappresentare sia la comunicazione tra esperti sia la comunicazione tra esperti e non esperti.

A tal scopo è stato adottato il framework teorico degli studi sugli usi dell'italiano istituzionale e la metodologia di analisi della terminologia istituzionale elaborata da Daniela Vellutino (Vellutino, 2014, 2015, 2018; Vellutino et al., 2012). Questo framework teorico si basa sull'individuazione degli obblighi informativi a cui si correlano i contesti comunicativi e i relativi tipi di testo istituzionali. Da qui scaturisce un modello di classificazione dei testi istituzionali: il modello CPI - Comunicazione Pubblica e Informazione istituzionale (Vellutino, 2018; Vellutino et al., 2012), che definisce i diversi tipi di testo istituzionale, per ogni obbligo informativo e ogni finalità delle attività di informazione e comunicazione delle pubbliche amministrazioni, come stabiliti dalla Legge 150/2000 [1] che le disciplina.

Esaminando i tipi di testo individuati dal modello CPI, emerge che atti legislativi, atti amministrativi e testi tecnico-operativi rappresentano la comunicazione tra esperti. Invece, i testi per l'ac-

[1] Legge 150 del 2000 sulla "Disciplina delle attività di informazione e di comunicazione delle pubbliche amministrazioni"

countability, i testi informativi e i testi delle campagne di comunicazione istituzionale rappresentano la comunicazione tra esperti e non esperti.

Ai fini del presente lavoro, è particolarmente importante ricostruire la filiera documentale dell'informazione istituzionale, che nasce nei documenti fonte (normativi e amministrativi) e viene ripresa nei testi informativi (brochure, siti web, avvisi pubblici, ecc.). Ad esempio, l'informazione che riguarda i giorni di raccolta dei rifiuti differenziati viene stabilita ufficialmente nelle ordinanze, che servono da testo fonte per poi redigere i calendari di raccolta differenziata, le brochure distribuite ai cittadini e le news sul sito web istituzionale.

Lungo questa filiera, la terminologia subisce una trasformazione dovuta a fattori di scopo della comunicazione, di registro comunicativo e di mezzo. È quindi fondamentale rappresentare, in un corpus, tutte le tipologie testuali che ne fanno parte.

Perciò uno degli obiettivi principali del presente lavoro è quello di sviluppare, a partire da questo framework teorico e metodologico, due corpora specialistici: il corpus ItaIst-DdAC_GRU e il corpus ItaIst-TermRifiuti. Entrambi i corpora sono descritti nelle sezioni che seguono.

4.2 ItaIst-DdAC_GRU

ItaIst-DdAC_GRU è un corpus specialistico primario composto da testi istituzionali relativi al dominio della gestione dei rifiuti urbani. Il corpus nasce dagli studi sull'italiano istituzionale (Vellutino, 2014, 2018; Vellutino et al., 2012) e si sviluppa ulteriormente nell'ambito del presente lavoro e del progetto PRIN 2020 "VerbACxSS: su verbi analitici, complessità, verbi sintetici e semplificazione. Per l'accessibilità" (2020, Prot. 020BJKB9M)[2]. Gli scopi che hanno portato allo sviluppo del corpus sono tre:

- Studiare la terminologia istituzionale della gestione dei rifiuti urbani;
- Delineare il profilo linguistico-terminologico della comunicatzione istituzionale tra esperti e quella tra esperti e non esperti ed evidenziarne similarità e differenze allo scopo di rintracciare i problemi di leggibilità;
- Insegnare ai futuri comunicatori pubblici a classificare e repertoriare i testi istituzionali e a riconoscerne le specificità.

4.2.1 Raccolta e metadatazione dei testi

ItaIst-DdAC_GRU è un corpus di monitoraggio che viene aggiornato annualmente tramite i *project work* degli studenti del corso "Comunicazione pubblica e linguaggi istituzionali" nel corso di laurea magistrale "Corporate Communication e Media" dell'Università degli Studi di Salerno[3].

In questo progetto, ogni studente monitora il sito web del proprio comune di residenza per raccogliere informazioni e documenti relativi alla gestione dei rifiuti urbani. Inoltre, quando i documenti

[2]Progetto di Rilevante Interesse Nazionale – PRIN 2020 "VerbACxSS: su verbi analitici, complessità, verbi sintetici e semplificazione. Per l'accessibilità"(2020, Prot. 2020BJKB9M), finanziato dal Ministero dell'Università e della Ricerca. Coinvolge l'Università Roma Tre (PI: Anna Pompei), l'Università del Molise (coordinamento dell'unità di ricerca: Giuliana Fiorentino) e l'Università di Salerno (coordinamento dell'unità di ricerca: Daniela Vellutino)

[3]La titolare del corso "Comunicazione pubblica e linguaggi istituzionali" è la Prof.ssa Daniela Vellutino, che ha definito e descritto nei materiali didattici il metodo glottodidattico e gli strumenti di raccolta e metadatazione dei testi qui descritti e utilizzati per il presente lavoro di ricerca.

la cui pubblicazione è obbligatoria non sono presenti,[4] gli studenti li richiedono esercitando il diritto di accesso civico semplice (Art. 5,Art. 40 D.Lgs. 33/2013).

Dopo aver raccolto i documenti, gli studenti li archiviano secondo il modello di classificazione dei testi istituzionali per le attività comunicazione pubblica e informazione istituzionale "CPI" e compilano, per ogni documento, una scheda di metadatazione che si basa sul modello "CPI" e contiene le seguenti informazioni:

- Nome originale del file (`original_filename`).
- Link alla pagina web su cui è presente il documento (`page_link`).
- Link per il download del documento (`download_link`).
- Ente che ha prodotto il documento (`author`).
- Tipo di documento (`document_type`).
- Titolo (`title`).
- Data di produzione del documento (`production_date`).
- Data di download del documento (`download_date`).
- Formato (`original_format`).
- Obblighi informativi e finalità (`itaist_purpose`).
- Funzione pragmatica (`itaist_pragmatic_function`).
- Contesto comunicativo (`itaist_communicative_context`).
- Tipo di testo istituzionale (`itaist_text_type`).
- Vincolo interpretativo (`itaist_interpretative_bond`).

Allo stato attuale, la raccolta di documenti è stata effettuata per tre anni consecutivi, con modalità leggermente diverse da un anno all'altro. Tuttavia, solo la prima versione del corpus con i dati raccolti durante il primo anno è stata pubblicata.

Costruzione (anno accademico 2022/23). La prima versione del corpus è il risultato di un processo esplorativo. Pertanto, la quantità di testi istituzionali raccolti non è stata definita a priori né è stato previsto un loro bilanciamento (Vellutino & Cirillo, 2024). Il risultato è un corpus composto principalmente da atti amministrativi (45%) e testi informativi (42%). I testi tecnico-operativi e per l'accountability, invece, sono minoritari (rispettivamente 9% e 4%).

[4] I documenti richiesti con la richiesta di accesso civico sono: Piano Economico Finanziario (PEF) del servizio di raccolta differenziata; documentazione dell'appalto della gestione dei rifiuti urbani; Modello unico Dichiarazione Ambientale (MUD); atti amministrativi sulla gestione dei rifiuti; piano di comunicazione della raccolta differenziata.

1° aggiornamento (anno accademico 2023/24). Per il primo aggiornamento del corpus, gli studenti hanno raccolto ciascuno 12 documenti prodotti nell'intervallo temporale 2021-2023, bilanciati per funzione (4 con funzione prescrittiva; 4 con funzione strumentale-regolativa e 4 con funzione informativa). È comunque stata prevista una certa flessibilità nel caso in cui non fosse possibile trovare, sul sito di un comune, abbastanza testi per ogni funzione.

2° aggiornamento (anno accademico 2024/25). Per il secondo aggiornamento, il corpus è stato integrato con post sui social media Facebook e Instagram. Nello specifico, gli studenti hanno raccolto ciascuno 9 documenti (3 con funzione prescrittiva; 3 con funzione strumentale-regolativa e 3 con funzione informativa) e 12 post sui social media così divisi:

- 3 post Facebook pubblicati dal comune;
- 3 post Facebook pubblicati dal gestore del servizio;
- 3 post Instagram pubblicati dal comune;
- 3 post Instagram pubblicati dal gestore del servizio.

Anche in questo caso è stata prevista flessibilità qualora non fossero presenti, sul sito di un comune, abbastanza testi per ogni funzione o sufficienti post sui social media.

4.2.2 Estrazione del testo e preprocessing

Per estrarre il testo dai documenti abbiamo sviluppato uno script Python in grado di trattare quattro tipologie differenti di file. Lo script tratta i PDF testuali attraverso la libreria *pdfplumber* e i documenti in formato DOCX attraverso la libreria *textract*. Più complesso è il trattamento di immagini e scansioni. Per estrarre il testo da tali documenti è necessario un software OCR (*Optical Character Recognition*). A questo scopo abbiamo optato per il software open source *Tesseract*. Infine, lo scraping dei comunicati stampa e delle news dei siti web istituzionali è stato effettuato tramite la libreria *Beautiful Soup*.

È molto comune che le tabelle vengano rimosse nel processo di costruzione di un corpus. Nel nostro caso abbiamo scelto invece di conservarle in quanto esse rappresentano una fonte preziosa per estrarre la terminologia specialistica.

Una volta estratto il testo dai documenti, si è poi reso necessario modificarne la formattazione per ricostruire le frasi. A questo scopo, la scelta è ricaduta sul modulo *sentence-splitter* sviluppato da Koehn (2005) per il trattamento del corpus Europarl. Infine, abbiamo ottenuto l'annotazione linguistica delle frasi (*part-of-speech tagging*, lemmatizzazione e *depency parsing*) grazie al parser *stanza* (Qi et al., 2020).

4.2.3 Composizione del corpus

La prima versione del corpus ItaIst-DdAC_GRU, descritta in Vellutino e Cirillo (2024), è composta da 306 testi istituzionali provenienti da 27 comuni della Campania. I testi coprono un periodo temporale che va dal 2005 al 2023 e riguardano il dominio terminologico della gestione dei rifiuti urbani. La maggior parte di essi appartiene alle tipologie testuali degli atti amministrativi (140) e dei testi informativi (127). Altre tipologie testuali repertoriate sono i testi tecnico-operativi (26)

e, meno rappresentati, i testi per l'accountability (13). La composizione del corpus è illustrata in Tabella 4.1.

Le principali tipologie di atti amministrativi presenti nel corpus sono i testi dei linguaggi istituzionali speciali, in particolare, gli atti amministrativi: delibere, moduli, determine, ordinanze, regolamenti, bandi, disciplinari di gara, capitolati d'appalto e carte dei servizi. Appartengono ai tipi di testo dei linguaggi istituzionali speciali anche i testi tecnico-operativi relativi alla gestione dei rifiuti quali il Modello Unico di Dichiarazione Ambientale (MUD) e il Piano Economico Finanziario (PEF) del servizio di gestione dei rifiuti urbani.

È da rilevare che alcuni testi per l'accountability sono in forma di tabelle di dati relativi alla raccolta differenziata, report e testi contabili quali i bilanci d'esercizio. Tra i testi istituzionali informativi troviamo invece gli avvisi pubblici che, dal punto di vista pragmalinguistico, rappresentano una forma testuale per definizione amministrativa e per funzione informativa (Vellutino, 2018).

I testi informativi presenti nel corpus sono per la maggior parte calendari e guide per il conferimento dei rifiuti e testi istituzionali mediali quali comunicati stampa e news per il sito web istituzionale.

Per quanto riguarda le dimensioni, il corpus contiene 1.351.921 token. È da rilevare che gli atti amministrativi sono, in media, molto più lunghi delle altre tipologie testuali (7.293 token per documento). Seguono i testi tecnico-operativi (7.066 token per documento), i testi per l'accountability (1.702 token per documento) e, infine, i testi informativi (1.212 token per documento) che sono i più brevi.

4.3 ItaIst-TermRifiuti

Una porzione del corpus ItaIst-DdAC_GRU è stata selezionata per l'annotazione manuale dei termini, dando vita al corpus ItaIst-TermRifiuti. ItaIst-TermRifiuti è un corpus specialistico secondario composto da testi istituzionali relativi al tema "gestione dei rifiuti urbani". I testi, estratti da ItaIst-DdAC_GRU, sono annotati da quattro annotatori che hanno identificato manualmente i termini. ItaIst-TermRifiuti nasce con 3 obiettivi principali:

- Studiare la terminologia istituzionale del dominio della gestione dei rifiuti urbani.
- Fornire un *gold-standard* per testare e addestrare algoritmi di EAT specializzati nell'isolare un determinato strato terminologico.
- Sviluppare una metodologia per la costruzione di corpora annotati che rispetti i principi *Findable, Accessible, Interoperable, Reusable* (FAIR) (Wilkinson et al., 2016) e il paradigma *Linguistic Linked Open Data* (LLOD).

Il campione che compone ItaIst-TermRifiuti comprende atti amministrativi e testi informativi, bilanciati per numero di *token* (Tabella 4.2). I documenti del campione sono stati selezionati attraverso la tecnica del campionamento casuale proporzionale semplice, ripuliti manualmente e suddivisi in paragrafi. Il corpus include integralmente i documenti che contengono meno di 1.000 *token* e contiene paragrafi casuali estratti dai documenti più lunghi.

Infine, i termini sono stati annotati tramite il software *doccano* (Nakayama et al., 2018) con le seguenti etichette di dominio:

Tipo di testo		**N° documenti**	**N° token**
atti amministrativi	delibere	44	334.393
	moduli	25	65.615
	ordinanze	21	99.354
	regolamenti	14	50.309
	determine	7	15.911
	bandi di gara	6	41.611
	carte dei servizi	5	11.447
	disciplinari di gara	4	3.415
	capitolati di appalto	4	3.802
	altri	10	31.372
	TOT	**140**	**657.229**
testi informativi	avvisi pubblici	35	89.644
	calendari di raccolta	32	156.878
	guide al conferimento	35	69.848
	comunicati stampa	14	40.971
	altro	11	127.933
	TOT	**127**	**485.274**
testi tecnico-operativi	MUD	11	48.628
	PEF	6	38.357
	altro	9	59.560
	TOT	**26**	**146.545**
testi per l'accountability	Dati sulla raccolta differenziata	7	24.427
	report	5	37.634
	bilanci di esercizio	1	8.120
	TOT	**13**	**62.873**
TOT		**306**	**1.351.921**

Tabella 4.1: Composizione del corpus ItaIst-DdAC_GRU

- Termine relativo alla gestione dei rifiuti (`waste management`);
- Termine relativo all'ambiente (`environment`);
- Termine relativo al diritto e all'amministrazione (`law`);
- Termine relativo al diritto dell'ambiente (`environmental law`);
- Termine relativo al diritto della gestione dei rifiuti (`waste management law`).

Ogni annotatore ha identificato, in media, 108 termini ogni 1000 *token*. Di questi termini, il 47% appartiene al dominio della gestione dei rifiuti e il 42% appartiene al dominio del diritto e dell'amministrazione. Le altre categorie sono residuali e insieme coprono il restante 11%.

In Figura 4.1 è illustrata la distribuzione dei termini nei diversi tipi di testo. Analizzando i dati, appare evidente come gli atti amministrativi contengano una quantità maggiore di termini del diritto e dell'amministrazione rispetto ai termini della gestione dei rifiuti. In particolare, i termini della gestione dei rifiuti sono residuali in delibere/determine (9%) e bandi/disciplinari di gara (3%). Invece, in ordinanze, carte dei servizi e regolamenti c'è un maggiore bilanciamento tra i due domini di conoscenza.

Per quanto riguarda i testi informativi, questi ultimi contengono in misura maggiore termini relativi alla gestione dei rifiuti. Nello specifico, calendari di raccolta e guide al conferimento contengono quasi esclusivamente termini della gestione dei rifiuti (rispettivamente 95% e 93%), mentre

Tipo di testo		N° documenti	N° token
atti amministrativi	ordinanze	6	4.259
	carte dei servizi	4	4.179
	bandi/disciplinari di gara	5	4.094
	delibere/determine	5	4.903
	regolamenti	5	4.216
	TOT	**25**	**21.651**
testi informativi	avvisi pubblici	12	5.049
	calendari	9	5.029
	guide	10	5.190
	news e comunicati stampa	13	4.414
	TOT	**44**	**19.682**
TOT		**69**	**41.333**

Tabella 4.2: Composizione del corpus ItaIst-TermRifiuti

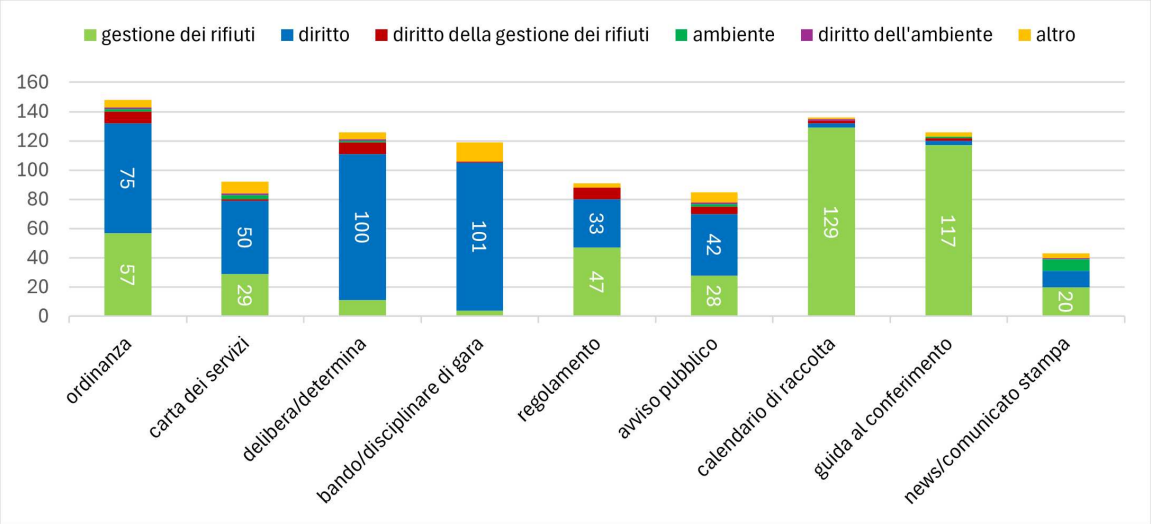


Figura 4.1: Numero medio di termini annotati ogni 1000 token per tipo di testo

gli avvisi pubblici, pur essendo testi informativi, contengono soprattutto termini del diritto (49%). Questo fenomeno conferma l'ipotesi che gli avvisi pubblici hanno finalità informativa, ma presentano la struttura tipica degli atti amministrativi (Vellutino, 2018). Infine, le news/comunicati stampa sono i testi che contengono meno termini.

4.3.1 Accordo tra annotatori

Il corpus ItaIst-TermRifiuti è stato annotato da quattro annotatori: l'autore di questo lavoro e tre volontari, di cui un dottorando di Linguistica dell'Università La Sapienza, collaboratore del progetto PRIN 2020 "VerbACxSS: su verbi analitici, complessità, verbi sintetici e semplificazione. Per l'accessibilità", una studentessa del corso di laurea in Corporate Communication e Media, che ha frequentato il corso "Comunicazione pubblica e linguaggi istituzionali"; e un'altra studentessa laureata triennale in Lettere Moderne e attuale corsista del corso di laurea binazionale Italia - Germania "Linguistica e Didattica dell'Italiano nel Contesto Internazionale" dell'Università di Salerno con l'Universität des Saarlandes. Gli annotatori hanno dunque diversa formazione linguistica.

Fase	Micro F1	Macro F1	Type F1
1° fase di addestramento	0,55	0,53	0,50
2° fase di addestramento	0,58	0,55	0,56
accordo finale	0,63	0,55	0,53
corpus completo	0,60	0,50	0,52

Tabella 4.3: *Inter-Annotator Agreement*

Dominio	**Micro F1**	**Type F1**
gestione dei rifiuti	0,65	0,54
diritto	0,54	0,50
ambiente	0,30	0,33
diritto della gestione dei rifiuti	0,33	0,33
diritto dell'ambiente	0,10	0,11
altro	0,17	0,19

Tabella 4.4: *Inter-Annotator Agreement* per dominio.

L'annotazione vera e propria è stata preceduta da una fase preliminare di addestramento degli annotatori e di discussione delle linee guida, come in Qasemizadeh e Schumann (2016), con la differenza che i diversi *round* di annotazione non sono stati effettuati sullo stesso campione, bensì su campioni diversi. In questo modo, si riduce il rischio che l'*Inter-Annotator Agreement* - IAA aumenti perché gli annotatori hanno preso accordi su specifiche unità lessicali presenti nel campione.

Durante la fase di addestramento, i quattro annotatori hanno effettuato due *round* di annotazione, seguiti ognuno da una riunione di discussione che ha portato all'aggiornamento delle linee guida. In questa fase, gli annotatori hanno lavorato su due campioni composti ciascuno da 20 paragrafi di lunghezza compresa tra i 200 e 300 caratteri. Per valutare se le performance degli annotatori sono effettivamente migliorate durante l'addestramento, al termine dell'annotazione, l'accordo è stato calcolato su un terzo campione, di composizione analoga ai precedenti.

Nello specifico, abbiamo calcolato l'IAA con tre tecniche diverse. La prima tecnica, *Macro F1*, descritta da Qasemizadeh e Schumann (2016), consiste nel calcolare l'accordo per ogni documento e poi farne una media. Invece, nella seconda tecnica denominata *Micro F1* (Verborgh et al., 2018) l'accordo è calcolato su tutto il corpus. Infine, la terza metrica, *Type F1*, considera solo i *type* dei termini individuati, ovvero conta ogni termine unico una sola volta, indipendentemente dalla sua frequenza nel corpus.

La Tabella 4.3 mostra l'accordo medio durante le fasi di addestramento e alla fine dell'annotazione. Possiamo osservare come l'accordo sia leggermente aumentato durante le due fasi di addestramento e discussione delle linee guida (*Micro F1*: +0,08; *Macro F1*: +0,02; *Type F1*: +0,03), a conferma dell'utilità di queste fasi.

Infine, per valutare se il dominio terminologico e il tipo di testo influenzino l'accordo, abbiamo calcolato l'IAA relativa a questi due parametri. La Tabella 4.4 illustra l'accordo medio per dominio[5] mentre la Tabella 4.5 mostra l'accordo medio per tipo di testo.

Possiamo notare che l'IAA relativo al dominio della gestione dei rifiuti (*Micro F1*: 0,65; *Type F1*: 0,54) è più alto di quello relativo ai termini del diritto che, benché inferiore, è comunque discreto

[5]In questo caso, L'IAA *Macro F1* non è stato calcolato perché risente molto della differenza tra le quantità di termini annotati per ogni dominio.

Tipo di testo		Micro F1	Macro F1	Type F1
atti amministrativi	ordinanza	0,66	0,58	0,58
	carta dei servizi	0,50	0,49	0,51
	delibera/determina	0,57	0,58	0,54
	bando/disciplinare di gara	0,60	0,59	0,56
	regolamento	0,49	0,49	0,51
	tutti gli atti amministrativi	**0,57**	**0,56**	**0,53**
testi informativi	avviso pubblico	0,50	0.50	0,47
	calendario di raccolta	0,69	0,64	0,61
	guida al conferimento	0,72	0,65	0,58
	news/comunicato stampa	0,41	0,47	0,41
	tutti i testi informativi	**0,63**	**0,59**	**0,51**

Tabella 4.5: *Inter-Annotator Agreement* per tipo di testo

(*Micro F1*: 0,54; *Type F1*: 0,50). Al contrario, l'accordo relativo ai domini residuali: (diritto della gestione dei rifiuti, ambiente, diritto dell'ambiente e altro) è decisamente basso.

Questo risultato suggerisce che, più è chiaro il dominio di conoscenza oggetto di analisi, più è facile decidere se un'unità lessicale è un termine o meno. Per quanto riguarda l'accordo per tipo di testo, rileviamo un valore leggermente maggiore per i testi informativi che potrebbe essere dovuto alla maggiore presenza di termini relativi alla gestione dei rifiuti in questi tipi di testo.

4.4 Formati di rappresentazione e modalità di interrogazione

Per aderire il più possibile ai principi FAIR (Wilkinson et al., 2016), i corpora ItaIst-DdAC_GRU e ItaIst-TermRifiuti sono distribuiti su Zenodo[6] (Cirillo et al., 2025) con licenza CC-BY 4.0 in formato RDF e sono descritti tramite il paradigma dei LLOD (Cimiano et al., 2020), ovvero tramite le ontologie *NLP Interchange Format* (NIF) (Hellmann et al., 2013), *Portable Linguistic Annotation with OWL* (POWLA) (Chiarcos, 2012b) e *Ontologies of Linguistic Annotation* (OLiA) (Chiarcos, 2012a).

Inoltre, abbiamo messo a disposizione un *endpoint* SPARQL[7] grazie al quale è possibile interrogare i corpora. Il punto di forza dell'*endpoint* SPARQL risiede nel fatto che permette di eseguire numerosi tipi di *query*, anche complesse e articolate, e di ottenere l'output in diversi formati, sia *human-readable* che *machine-readable*.

Di seguito sono riportate alcune query che esemplificano le potenzialità di interrogazione dei corpora. La prima *query*, illustrata in Figura 6.3 restituisce tutte le frasi in cui è presente una data forma di parola (che nell'esempio è *rifiuti*). Invece, la query in Figura 6.4 restituisce tutte le forme di un dato lemma (che nell'esempio è *conferire*) e le frasi in cui appare.

Ricercare una sequenza poliremática richiede invece una *query* leggermente più complessa, che prevede l'utilizzo del comando `FILTER`. In Figura 6.5 è mostrata una query che restituisce tutte le frasi che contengono la sequenza di lemmi *centro*, *di*, e *raccolta*.

Inoltre, nel corpus ItaIst-TermRifiuti, è possibile eseguire ulteriori interrogazioni che riguardano l'annotazione. Ad esempio, con la query in Figura 4.5 è possibile ritrovare tutti i termini appartenenti al dominio terminologico della gestione dei rifiuti, individuati da un determinato annotatore.

[6] https://doi.org/10.5281/zenodo.15173712
[7] http://193.205.161.60:8890/sparql/ (Ultima rilevazione: 10/04/2025)

```
SELECT ?sentence
WHERE {
    ?word nif:anchorOf "rifiuti"@it .
    ?sent nif:word ?word;
        nif:anchorOf ?sentence .
}
```

Figura 4.2: Ricerca per forma di parola.

```
SELECT ?form ?sentence
WHERE {
    ?word nif:lemma "conferire"@it;
        nif:anchorOf ?form .
    ?sent nif:word ?word;
        nif:anchorOf ?sentence .
}
```

Figura 4.3: Ricerca per lemma.

SPARQL permette anche di realizzare query molto complesse, come la query in Figura 4.6, che mostra tutti i termini del diritto e dell'amministrazione individuati da un dato annotatore in ogni frase del corpus.

Infine, grazie al collegamento tra ItaIst-TermRifiuti è la banca dati terminologica ItaIst-WasteLexicon, è possibile anche effettuare ricerche congiunte tra le due risorse (cfr. Capitolo 6).

```
SELECT ?form1 ?form2 ?form3 ?sentence
WHERE {
    ?word1 nif:lemma "centro"@it ;
        nif:anchorOf ?form1 ;
        nif:nextWord ?word2 .
    ?word2 nif:nextWord ?word3 ;
        nif:lemma ?lemma2 ;
        nif:anchorOf ?form2 .
    ?word3 nif:lemma ?lemma3;
        nif:anchorOf ?form3 .
    ?sent nif:word ?word1 ;
    nif:anchorOf ?sentence .
    FILTER(?lemma2 = "di"@it && ?lemma3 = "raccolta"@it)
}
```

Figura 4.4: Ricerca per sequenza di lemmi

```
SELECT DISTINCT ?string
    WHERE {
    ?annotation itsrdf:termAnnotatorsRef termrifiuti:annotator_1 ;
       termrifiuti:hasLabel "Term (waste)"@en ;
       powla:string ?string ;
       powla:hasChild*
    }
```

Figura 4.5: Ricerca dei termini della gestione dei rifiuti annotati da un dato annotatore

```
SELECT DISTINCT (GROUP_CONCAT (DISTINCT ?string; separator=', ') AS ?terms) ?sentence
    WHERE {
    ?annotation itsrdf:termAnnotatorsRef termrifiuti:annotator_1 ;
       termrifiuti:hasLabel "Term (law/administration)"@en ;
       powla:string ?string ;
       powla:hasChild{2}/^nif:word/nif:anchorOf ?sentence.
    }
    GROUP BY ?sentence
```

Figura 4.6: Ricerca per frase di tutti i termini del diritto e dell'amministrazione individuati da un dato annotatore

Capitolo 5

Domain Concept Relatedness

Contenuti

5.1 Isolare la terminologia di dominio 51
5.2 Key Concept Relatedness 52
5.3 Domain Concept Relatedness 52
5.4 Sperimentazione 53
5.4.1 Corpora 53
5.4.2 Dataset 54
5.4.3 Metodologia di valutazione 54
5.4.4 Risultati 55

5.1 Isolare la terminologia di dominio

Un tipico algoritmo di Estrazione Automatica di Termini (EAT) estrae tutti i termini presenti in un corpus senza considerare il dominio terminologico a cui tali termini appartengono. Anche se quest'approccio risulta ragionevole nella maggior parte dei casi, ci sono dei contesti in cui è necessario distinguere i termini dei diversi domini. A questo proposito, Lenci et al. (2009) e Bonin et al. (2010) affermano che quando l'EAT è applicata ai documenti legislativi, diviene cruciale isolare i termini che appartengono al settore oggetto della normativa dai termini del diritto. Tuttavia, allo stato dell'arte, solo pochi algoritmi di EAT affrontano questo specifico compito (cfr. Capitolo 3).

Nei linguaggi istituzionali, isolare la terminologia di uno specifico dominio è un compito fondamentale perché il lessico dei linguaggi istituzionali proviene da differenti domini di conoscenza, tra i quali il diritto, l'amministrazione, l'economia, la finanza e tutti i settori di cui si occupano le diverse istituzioni (Vellutino, 2018).

Per questo motivo, abbiamo sviluppato e testato una nuova tecnica di EAT in grado di isolare uno specifico Strato Terminologico (ST) (cfr. Capitolo 2) in contesti lessicalmente complessi come quello dei linguaggi istituzionali. Questa tecnica, che abbiamo chiamato *Domain Concept Relatedness* (DCR), si basa sulla semantica distribuzionale ed è un'evoluzione di un algoritmo preesistente, denominato *Key Concept Relatedness* (KCR) (Astrakhantsev, 2018). L'obiettivo della DCR è quello di isolare, in un corpus specialistico, un determinato ST.

5.2 Key Concept Relatedness

La KCR (Astrakhantsev, 2018) è una tecnica di EAT basata sull'assunto che i termini di un dato dominio sono correlati semanticamente a concetti chiave già conosciuti appartenenti allo stesso dominio. Per misurare la similarità semantica (*relatedness*), la KCR calcola la similarità di coseno tra *embedding* di concetti. Nell'ultima versione della KCR, gli *embedding* di concetti sono generati a partire da *hyperlink* delle pagine Wikipedia.

Nello specifico, i file *dump* di Wikipedia sono stati trattati rimuovendo i segni di marcatura e sostituendo ogni *hyperlink* con un token speciale, che rappresenta un concetto. A questo punto, l'*embeding* di ogni concetto è generato tramite l'algoritmo *word2vec* (Mikolov et al., 2013).

Successivamente, per ottenere i concetti chiave, la KCR utilizza una versione semplificata dell'algoritmo di estrazione di *keyword* KP-Miner (El-Beltagy & Rafea, 2010) che estrae automaticamente i concetti chiave da ogni documento del corpus. Infine, per ogni termine `t` che possiede un *embedding*, la KCR calcola la sua correlazione semantica con un insieme di `k` concetti chiave, selezionati attraverso l'algoritmo *k Nearest Neighbor* con solo esempi positivi. La formula per calcolare la correlazione semantica è illustrata nell'Equazione 5.1.

$$relatedness(t, C_d) = \frac{1}{k} \sum_{i=1}^{k} cos(v_t, v_i) \tag{5.1}$$

dove C_d è un insieme di concetti chiave ordinato per similarità di coseno con il candidato termine t, k è un parametro dell'algoritmo kNN; v_t e v_i sono rispettivamente le rappresentazioni vettoriali del termine t e del concetto chiave i. Se un candidato termine non ha una rappresentazione vettoriale, la *relatedness* corrispondente a quel termine è uguale a 0.

5.3 Domain Concept Relatedness

L'algoritmo della DCR prevede tre fasi: l'estrazione dei candidati, la generazione degli *embedding* di concetto e il calcolo della *relatedness*.

Nella fase di estrazione dei candidati, la DCR impiega dei *pattern* morfosinattici basati sullo strumento *RegexpParser* della libreria *nltk*. Dato che il contributo più importante della DCR risiede nel meccanismo di generazione degli *embedding*, abbiamo deciso di utilizzare una tecnica di estrazione dei candidati molto semplice. La DCR, quindi, estrae solo nomi e sintagmi nominali.

Come già accennato, l'obiettivo della DCR è quello di estrarre solo i termini che appartengono al dominio oggetto di analisi, ignorando altri termini che potrebbero apparire nel corpus specialistico. La KCR, così come concepita, non è adatta a isolare uno strato terminologico. Inoltre, è inadatta a trattare quei domini che hanno una scarsa copertura in Wikipedia.

Per affrontare la prima questione, abbiamo reso la DCR semi-supervisionata: i concetti chiave sono presi da una risorsa terminologica esistente o inseriti dall'utente. Questa modifica assicura che tutti i concetti chiave appartengano al dominio oggetto di analisi e permette alla DCR di isolare un dato ST. Inoltre, per superare il problema della copertura di Wikipedia, la DCR impiega una tecnica diversa per produrre gli *embedding* di concetti. Mentre la KCR li produce a partire dai file *dump* di Wikipedia, la DCR produce gli *embedding* di concetti direttamente dal corpus specialistico attraverso l'algoritmo *Alacarte* (Khodak et al., 2018). Questa tecnica ha due vantaggi: risolve il problema della copertura di Wikipedia e assicura che ogni candidato termine abbia la rispettiva rappresentazione

Corpus	Dominio	Documenti	Frasi	Token
principale	gestione dei rifiuti	148	15.463	630.456
contrastivo	trasporti	60	8.239	331.061
	sanità	28	3.271	130.708
	sicurezza	15	1.959	71.005
	agricoltura	56	1.807	68.859
	Totale	159	15.276	601.633

Tabella 5.1: Composizione dei corpora.

vettoriale. Il vantaggio di *Alacarte*, rispetto ai modelli di *word embedding* tradizionali, è la sua abilità di produrre le rappresentazioni in tempi brevi e su corpora relativamente piccoli. Inoltre, per quanto riguarda le parole poco frequenti, *Alacarte* produce rappresentazioni più affidabili dei modelli tradizionali. Per costruire la matrice di trasformazione, *Alacarte* necessità di un set di *word embedding* preaddestrato e del corpus usato per l'addestramento. A questo scopo, abbiamo addestrato un modello *word2vec* (Mikolov et al., 2013) sul corpus *Paisà* (Lyding et al., 2014).

Una volta generati gli embedding dei concetti, la *relatedness* è calcolata come nella KCR (Equazione 5.1).

5.4 Sperimentazione

Per testare la DCR, abbiamo deciso di compararla a due algoritmi di EAT proposti in letteratura. Il primo è il metodo *simple maths* (Kilgarriff, 2009), utilizzato dal tool di EAT presente nel software *Sketch Engine*. Questo metodo compara la frequenza dei candidati termini nel corpus specialistico con la loro frequenza in un corpus generale di riferimento. *Simple maths* non è pensato per isolare uno strato terminologico, perciò dovrebbe fornire un indizio sulla distribuzione dei diversi strati terminologici nel corpus. Inoltre lo abbiamo preferito a tecniche simili in quanto *Sketch Engine* è uno strumento molto diffuso tra i terminologi.

Il secondo metodo è l'approccio contrastivo (Bonin et al., 2010), una tecnica che è già stata utilizzata per isolare la terminologia del settore regolamentato in un corpus di documenti legislativi.

Nella valutazione, i concetti chiave forniti in *input* alla DCR sono estratti dal glossario /textitZerosprechi[1] che contiene 58 termini e rappresenta un esempio di glossario online di facile reperibilità.

5.4.1 Corpora

Per l'esperimento, abbiamo appositamente costruito due corpora (di circa 600.000 token ognuno): un corpus principale e un corpus contrastivo. Entrambi i corpora sono composti esclusivamente da direttive e regolamenti europei. Il corpus principale riguarda il dominio della gestione dei rifiuti, mentre il corpus contrastivo, necessario a testare l'approccio contrastivo, copre quattro argomenti diversi (agricoltura e pesca, sanità pubblica, sicurezza sul lavoro e trasporti). La struttura dei corpora è riassunta in Tabella 5.1.

Per costruire il corpus principale, abbiamo raccolto tutte le direttive e i regolamenti relativi ad un'entrata Eurovoc che riguarda la gestione dei rifiuti o ad uno dei suoi iponimi (Vellutino et al., 2016). Invece, le direttive e i regolamenti che compongono il corpus contrastivo riguardano materie

[1] https://www.zerosprechi.eu/index.php/glossario

differenti che, secondo noi, sono abbastanza varie da permettere all'approccio contrastivo di catturare i termini che caratterizzano direttive e regolamenti in generale.

5.4.2 Dataset

Il *dataset* di valutazione è stato costruito selezionando i primi 200 termini monorematici e i primi 200 termini polirematici estratti dal corpus principale attraverso la DCR, il metodo *simple maths* (Kilgarriff, 2009) e l'approccio contrastivo (Bonin et al., 2010). Il *dataset* finale contiene, in totale, 1.039 elementi lessicali unici.

Gli elementi lessicali sono stati valutati da tre annotatori: l'autore di questo lavoro e due studenti del corso di laurea magistrale in Corporate Communication e Media con una discreta conoscenza del dominio. Ogni annotatore ha ricevuto la lista degli elementi lessicali ordinati casualmente, le linee guida di annotazione e il corpus principale.

Per ogni elemento lessicale gli annotatori hanno controllato il suo uso nel corpus e hanno fornito due giudizi. Innanzitutto hanno deciso se l'elemento è un termine. Poi hanno specificato il dominio a cui appartiene tra cinque possibili: diritto (`LAW`); diritto della gestione dei rifiuti (`WASTE LAW`); gestione dei rifiuti (`WASTE`); altri argomenti legati alla gestione dei rifiuti (`WASTE REL`) e altri domini (`OTHER`).

In generale l'*Inter-Annotator Agreement* riguardante l'identificazione dei termini è moderato (k di Fleiss = 0.54) e l'accordo sui domini è leggermente inferiore (k di Fleiss = 0,47). Ciononostante, i risultati sono in linea con la letteratura: l'accordo tende ad essere basso perché non ci sono confini chiari tra la terminologia e il resto del lessico (Rigouts Terryn et al., 2020).

Nella versione finale del *dataset*, un elemento è considerato un termine valido se almeno due annotatori lo hanno giudicato tale. Lo stesso vale per le etichette di dominio, se almeno due annotatori hanno assegnato un termine allo stesso dominio, quell'etichetta è stata conservata nel dataset finale. In alternativa, l'annotatore principale (l'autore di questo lavoro) ha deciso quale etichetta riportare.

5.4.3 Metodologia di valutazione

Per una valutazione accurata dell'EAT è necessario disporre di un corpus annotato manualmente. Al tempo dello studio un corpus italiano sul dominio gestione dei rifiuti non era disponibile. Per questo motivo, abbiamo deciso di effettuare la valutazione sulla lista di unità lessicali estratta da ogni strumento. Questo approccio, però, presenta lo svantaggio di misurare solo la *precisione* (ovvero la correttezza delle unità estratte) ma non il *richiamo* (ovvero la percentuale delle unità estratte sul totale).

Nello specifico, per misurare la *precisione* abbiamo calcolato due metriche: la *precision at k* - P@k e la *precisione* media - PM. P@k è uguale al numero di termini corretti sui primi `k` elementi lessicali estratti dall'algoritmo. Più elementi lessicali sono termini, più è alto P@k. In questo esperimento k = 200. PM, invece, prende in considerazione anche l'ordine dei candidati. Di conseguenza, più i termini corretti appaiono in cima alla lista, più PM è alta. PM è calcolata con la Formula 5.2,

$$PM = \sum_{n}(R_n - R_{n-1})P_n \tag{5.2}$$

dove P_n e R_n sono la *precisione* e il *richiamo* alla soglia n. Anche la PM è stata calcolata considerando solo i primi 200 termini estratti da ogni algoritmo.

Tipo	Algoritmo	Tutti		Rifiuti		Diritto	
		P@200	PM	P@200	PM	P@200	PM
monorematici	simple maths	0.40	0.52	0.23	0.33	0.18	0.23
	approccio contrastivo	0.47	0.55	0.31	0.37	0.17	0.21
	DCR	0.20	0.28	0.17	0.29	0.03	0.02
polirematici	simple maths	0.49	0.53	0.29	0.36	0.21	0.20
	approccio contrastivo	0.54	0.67	0.40	0.54	0.15	0.15
	DCR	0.43	0.62	0.37	0.62	0.06	0.05

Tabella 5.2: Risultati della valutazione.

Nell'esperimento abbiamo effettuato tre valutazioni. Nella prima valutazione, tutti i termini sono considerati corretti indipendentemente dal dominio a cui appartengono. Nella seconda valutazione, solo i termini relativi alla gestione dei rifiuti (`WASTE`, `WASTE LAW` e `WASTE REL`) sono considerati corretti. Nell'ultima valutazione, abbiamo considerato corretti solo i termini del diritto (`LAW`).

5.4.4 Risultati

I risultati della valutazione sono illustrati in Tabella 5.2. Analizzando i risultati del metodo *simple maths* emerge che il corpus contiene termini della gestione dei rifiuti e termini del diritto in percentuali simili.

Al contrario l'approccio contrastivo e la DCR sono più adatti ad isolare la terminologia della gestione dei rifiuti. Specialmente il secondo, che ha estratto una percentuale minima di termini del diritto. Inoltre, la DCR ha ottenuto la PM più alta nell'estrazione di termini polirematici della gestione dei rifiuti mantenendo al contempo una P@200 molto simile all'approccio contrastivo. Questo significa che la DCR e l'approccio contrastivo estraggono un numero simile di termini ma la DCR tende a metterli in posizioni più alte nella lista.

Lo stesso non si può dire nel contesto dei termini monorematici dove la DCR ha ottenuto punteggi inferiori sia rispetto all'approccio contrastivo che al metodo *simple maths*.

In generale possiamo concludere che la DCR funziona bene con i termini polirematici dove riesce ad isolare uno strato terminologico meglio del metodo *simple maths* e dell'approccio contrastivo. Al contrario, non è adeguata al trattamento di termini monorematici.

Capitolo 6

Base dati terminologica ItaIst-Waste Lexicon

Contenuti

6.1 Introduzione 57

6.2 Metodologia di costruzione 58

6.3 Dimensione morfosintattica 59

6.3.1 Struttura morfologica dei termini monorematici 59

6.3.2 Struttura sintattica dei termini polirematici 60

6.4 Dimensione concettuale 61

6.5 Dimensione semantico-comunicativa 62

6.5.1 Rilevanza terminologica 63

6.5.2 Risultati 64

6.6 Formati di rappresentazione e modalità di interrogazione 65

6.1 Introduzione

ItaIst-WasteLexicon è una base di dati terminologica che riguarda la terminologia istituzionale della gestione dei rifiuti urbani. Contiene 910 termini che designano circa 380 concetti. Ad ogni termine sono associate proprietà morfosintattiche, concettuali e semantico-comunicative. Le proprietà morfosintattiche riguardano le forme flesse, la parte del discorso e i sotto-termini; le proprietà concettuali riguardano la definizione e le relazioni gerarchiche tra concetti, definite tramite le relazioni *genere/specie* e *parte/tutto*. Infine, le proprietà semantico-comunicative riguardano il grado di specialismo (Serianni, 2007; Vellutino, 2018) e la distribuzione del termine nei sotto-corpora di ItaIst-DdAC_GRU e ItaIst-TermRifiuti.

In Appendice A - Esempio di scheda terminologica è mostrato un esempio di scheda terminologica, per il termine *isola ecologica*.

6.2 Metodologia di costruzione

Prima di iniziare lo sviluppo della banca dati terminologica, è stato necessario approfondire la conoscenza dei processi amministrativi connessi al dominio della gestione dei rifiuti urbani e apprenderne la terminologia.

A tale scopo, l'autore del presente lavoro ha svolto un tirocinio formativo della durata di sei mesi presso un'azienda operante nel campo della gestione dei rifiuti urbani, Microambiente s.r.l.,[1] che si occupa di supportare aziende e comuni nel tenere traccia dei flussi dei rifiuti e nell'adempiere agli obblighi normativi in materia, tramite l'offerta di servizi e lo sviluppo di appositi software.

Nello specifico, l'autore del presente lavoro, durante il tirocinio, ha avuto l'opportunità di confrontarsi con numerosi esperti di dominio e ha fatto esperienza delle principali attività amministrative connesse alla gestione dei rifiuti, tra cui:

- La compilazione dei Formulari di Identificazione del Rifiuto (FIR);
- La Vidimazione Virtuale del Formulario di Identificazione del Rifiuto (Vi.Vi.Fir);
- La redazione dei Piani Economico Finanziari (PEF) della gestione dei rifiuti urbani;
- La trasmissione dei Modelli Unici di Dichiarazione ambientale (MUD).

Grazie a queste attività, è stato possibile apprendere sul campo la terminologia tecnica e amministrativa relativa al dominio della gestione dei rifiuti urbani e selezionare con maggiore consapevolezza i termini da includere nella banca dati.

Per costruire ItaIst-WasteLexicon abbiamo adottato un approccio di tipo semasiologico che parte dal termine per individuarne il significato e quindi delineare il concetto. Nello specifico, dal corpus ItaIst-TermRifiuti, sono stati raccolti tutti i termini relativi al dominio *gestione dei rifiuti*, individuati da almeno uno degli annotatori, sono stati scartati i termini errati o giudicati non rilevanti per il dominio considerato e, dopo un'attenta analisi del significato e del contesto d'uso, ogni termine è stato assegnato al rispettivo concetto. In questo modo, quando due termini sono sinonimi, essi sono assegnati al medesimo concetto.

Considerando che il focus di ItaIst-WasteLexicon è la variazione terminologica, abbiamo scelto di includere nella banca dati anche lunghe perifrasi qualora queste ultime siano utilizzate sistematicamente oppure quando esprimono un concetto che ha anche altre denominazioni più concise. Ad esempio, abbiamo incluso il termine *rifiuti provenienti da luoghi adibiti ad uso di civile abitazione*, perché rappresenta una perifrasi del termine *rifiuti domestici*.

Successivamente, ogni concetto è stato definito seguendo una precisa gerarchia di fonti:

1. Normativa europea ed italiana;[2]
2. IATE;

[1] https://wmysir.com/chi-siamo/ (Ultima rilevazione: 17/05/2025).

[2] Le principali fonti normative europee sono le seguenti: *Direttiva 1996/62/CE*; *Direttiva 2006/66/CE*; *Direttiva 2008/98/CE*; *Direttiva 2012/19/UE*; *Direttiva 2019/904/UE*; *Regolamento 2012/1179/UE*; *Regolamento 2024/1157/UE*. Mentre le fonti normative italiane sono le seguenti: *Decreto Legislativo 152/2006*; *Decreto legislativo 95/1992*; *Decreto Ministeriale 127/2024*; *Decreto Ministeriale 182/2019*; *Decreto Ministeriale 20/04/2017*; *Decreto del Presidente della Repubblica 2003/252*; T*esto unico per la regolazione della qualità del servizio di gestione dei rifiuti urbani (TQRIF)*; *Metodo tariffario rifiuti (MTR-2)*.

3. Dizionari ed enciclopedie online;

4. Siti web.

Nel caso in cui, per un determinato concetto non sia stata trovata una definizione nella normativa e le definizioni trovate nelle altre fonti non siano state ritenute soddisfacenti, abbiamo redatto la definizione prendendo come modello altre definizioni di concetti simili. Ad esempio, il termine *rifiuto pericoloso* è definito nella normativa come

> **rifiuto pericoloso:** rifiuto che presenta una o più caratteristiche pericolose di cui all'allegato III della Direttiva 2008/98/CE.

. Di conseguenza il termine *rifiuto ingombrante pericoloso* che invece non ha una definizione nella normativa è stato definito come:

> **rifiuto ingombrante pericoloso:** *rifiuto ingombrante* che presenta una o più caratteristiche pericolose di cui all'allegato III della Direttiva 2008/98/CE.

ItaIst-WasteLexicon è stato realizzato con l'ausilio dell'editor di risorse terminologiche VocBench 3 (Stellato et al., 2020)

6.3 Dimensione morfosintattica

I termini presenti in ItaIst-TermRifiuti sono quasi esclusivamente nomi (95%). Inoltre sono presenti, in misura residuale, aggettivi (3%) e verbi (2%). La maggior parte dei termini sono unità polirematiche (78%), come ad esempio *centro di raccolta* e *accumulatore esausto*; abbiamo poi le unità monorematiche (15%), come ad esempio *rifiuto*, *bioplastica* e *indifferenziato*. Inoltre, ItaIst-WasteLexicon contiene anche, in misura residuale, acronimi (3%), come ad esempio *CCR* (*Centro Comunale di Raccolta*) e *PAP* (*Porta-A-Porta*); strutture miste (2%), come ad esempio *piccolo RAEE* e *metodo MTR*; simboli di sostanze chimiche e di formule (1%), come ad esempio *al - alluminio*, *t - tossico* e *kb - coefficiente proporzionale di produttività* . Infine, ci sono alcuni codici alfanumerici (1%), come i codici del catalogo europeo dei rifiuti (es. *20 01 38 - rifiuti di legno*) e i codici delle operazioni di recupero (es. *R2 - rigenerazione/recupero di solventi*).

6.3.1 Struttura morfologica dei termini monorematici

Analizzando più nel dettaglio i termini monorematici e la loro struttura morfologica, emerge che il 48% di essi sono parole derivate. La Figura 6.1 mostra la distribuzione dei morfemi derivazionali nei termini derivati di ItaIst-WasteLexicon.

Di questi termini derivati, il 79% sono forme suffissate e il restante 21% sono forme prefissate. Tra i meccanismi di derivazione più frequenti troviamo la suffissazione zero, tipica dell'italiano istituzionale, che produce nomi deverbali come *deposito*, da *depositare*; *riciclo*, da riciclare, e *recupero*, da *recuperare*. Invece, la maggior parte degli aggettivi derivati presenta il suffisso *-bile*. Tra di essi troviamo aggettivi come *conferibile*, da *conferire*; *differenziabile* da *differenziare* e *biodegradabile*, da *biodegradare*. Seguono in frequenza i participi passati con funzione aggettivale e nominale, tra cui troviamo i termini *conferito*, *raccolta* e *riciclato*. Tra i prefissi, il prefisso *eco-* è il più frequente ed è presente in termini quali *ecocalendario*, *ecocentro* ed *ecovigilanza*. Abbiamo poi i suffissi *-mento*

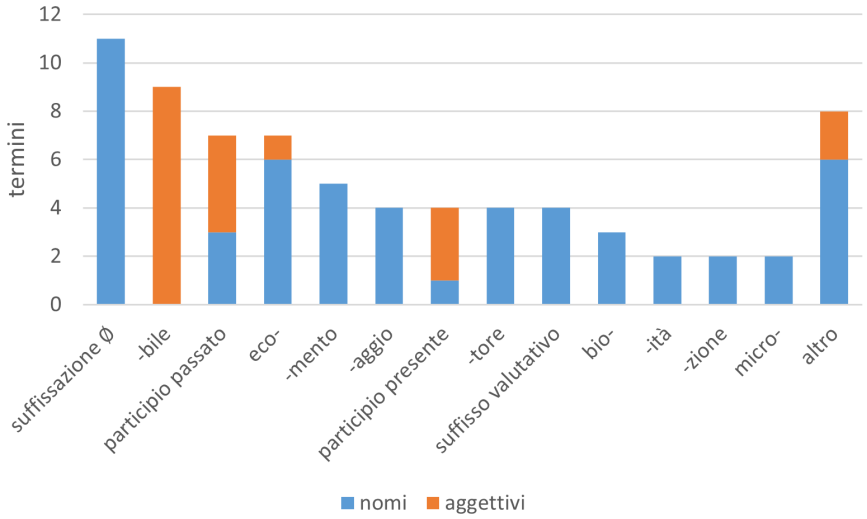


Figura 6.1: Morfemi derivazionali e nei termini di ItaIst-WasteLexicon

e *-aggio* che producono forme come *conferimento*, da *conferire* e *compostaggio* da *compostare*. Interessanti sono anche i termini derivati per alterazione, che presentano un suffisso valutativo, come *cestino* da *cesto*, *secchiello* da *secchio* e *cassonetto* da *cassone*.

Analizzando le basi da cui derivano i termini, emerge che la maggior parte dei termini derivati sono deverbali (48 termini), ci sono alcuni termini denominali (18 termini) e solo 2 termini sono deaggettivali. Le basi più frequenti sono i verbi *conferire*, *differenziare* e *riciclare*, da cui derivano 4 termini ciascuno.

La grande concentrazione di nomi deverbali nella terminologia istituzionale della gestione dei rifiuti urbani supporta l'ipotesi che l'italiano istituzionale faccia largo uso di nominalizzazioni e di costruzioni a verbo supporto.

6.3.2 Struttura sintattica dei termini polirematici

I termini polirematici presenti in ItaIst-WasteLexicon presentano un'ampia varietà di strutture sintattiche, anche a causa della scelta metodologica di includere lunghe perifrasi nella banca dati. Nello specifico, le due strutture più frequenti sono *nome + aggettivo* (19% dei termini polirematici), che è la struttura di termini come *abito usato*, *contenitore stradale* e *frazione organica* e *nome + preposizione + nome* (11%), che produce termini come *calendario di conferimento*, *raccolta di prossimità* e *scarto di potatura*. Meno frequenti sono le costruzioni *nome + preposizione + determinante + nome* (6%), come *combustione all'aperto*, *gestione dei rifiuti* e *igiene del suolo*. Seguono i composti *nome + nome* (4%) tra cui, ad esempio, *ritiro rifiuti*, *scarrabile pressa* e *tassa rifiuti*. Interessanti sono anche i composti *nome + aggettivo + aggettivo* la cui frequenza (3%) è particolarmente alta considerando la peculiarità del costrutto. Alcuni esempi di termini con questa struttura sono *frazione umida organica*, *isola ecologica itinerante* e *rifiuto solido urbano*.

Analizzando la lunghezza dei termini polirematici, intesa come numero di parole,[3] emerge che

[3]Nel calcolare la lunghezza dei termini polirematici, abbiamo considerato le preposizioni articolate come due parole

la maggior parte di essi è composta da due (27,6%), tre (21,8%) o quattro parole (15,7%). Anche i termini composti da cinque parole hanno una discreta frequenza (11,6%). I termini da sei a dieci parole rappresentano il 19,7% e il restante 3,6% dei termini è composto da più di dieci parole. Il termine più lungo, che contiene ben 25 parole, è il seguente.

> *pile portatili, batterie e accumulatori al piombo derivanti dalla manutenzione di veicoli ad uso privato effettuata in proprio dalle utenze domestiche*

Tale sintagma nominale, sebbene sia particolarmente complesso, ha, secondo noi, valenza terminologica. Infatti, esso appare nella stessa forma, in numerosi regolamenti dei centri di raccolta ed indica una specifica tipologia di rifiuto che può essere conferita presso il centro di raccolta.

6.4 Dimensione concettuale

L'organizzazione concettuale di ItaIst-WasteLexicon, effettuata attraverso relazioni gerarchiche del tipo *genere/specie* e *parte/tutto* ha prodotto una struttura concettuale imperniata su 5 categorie principali, sotto cui ricadono 195 concetti, e 8 categorie secondarie sotto cui ricadono 72 concetti. I restanti 101 concetti, invece, non ricadono sotto nessuna delle categorie precedenti[4].

Le categorie principali sono le seguenti:

- **Rifiuto** (59 concetti). È la categoria più numerosa. Raccoglie tutto ciò che, secondo la normativa, è classificabile come rifiuto. Ovvero "Qualsiasi sostanza od oggetto di cui il detentore si disfi o abbia l'intenzione o l'obbligo di disfarsi". Questa categoria include, tra gli altri, i concetti corrispondenti alle diverse frazioni merceologiche oggetto di raccolta differenziata come *carta e cartone*, *plastica*, *organico*, ecc.

- **Gestione dei rifiuti** (55 concetti) Include le diverse operazioni di gestione dei rifiuti come definite dalla normativa. Tali operazioni si dividono in *raccolta*, *trasporto*, *recupero* e *smaltimento*.

- **Attrezzatura** (35 concetti). Comprende tutti quei concetti che si riferiscono a strumenti, attrezzi e veicoli necessari al trattamento dei rifiuti e alla raccolta differenziata. Tra questi strumenti troviamo, ad esempio, il *sacchetto*, il *mastello*, l'*autospazzatrice* e la cosiddetta *card*: la tessera che serve a identificare i cittadini-utenti quando si recano al centro di raccolta.

- **Documento** (25 concetti). Comprende i documenti relativi alla gestione dei rifiuti, sia amministrativi che informativi. In particolare, tra i documenti amministrativi troviamo l'*autorizzazione unica ambientale* (AUA), l'*avviso di pagamento* e la *carta dei servizi*, mentre tra i documenti informativi troviamo il *calendario di raccolta*, la *guida al conferimento* e l'*ecodizionario*: una lista di oggetti comuni che indica le corrette modalità di conferimento per ciascun oggetto.

- **Caratteristica** (21 concetti). Comprende tutti gli aggettivi che esprimono caratteristiche del rifiuto come, ad esempio, *biodegradabile*, *differenziabile* e *pericoloso*. È interessante notare

distinte.

[4]Si noti che le categorie sono esse stesse concetti.

come, nella maggior parte dei casi, per un aggettivo è presente anche il suo opposto (es. *biodegradabie/non biodegradabile*; *pericoloso/non pericoloso*, ecc.).

Le categorie secondarie, invece, sono le seguenti:

- **Tassa sui rifiuti** (11 concetti). Comprende i concetti relativi alle tasse italiane sui rifiuti urbani. Tra di esse troviamo sia le tasse che sono state abolite, come la *tassa per lo smaltimento dei rifiuti solidi urbani* (TARSU) e il *tributo comunale sui rifiuti e sui servizi* (TARES), sia l'attuale *tassa rifiuti* (TARI), che si divide in *tassa rifiuti presuntiva* e *tassa rifiuti puntuale* a seconda del metodo di calcolo.
- **Impianto di trattamento** (14 concetti). Include i concetti relativi ai diversi impianti di trattamento dei rifiuti, distinti in base alla fase del ciclo dei rifiuti in cui si collocano (es. *impianto intermedio*; *impianto finale*) o al tipo di operazione svolta (es. *centro di stoccaggio*; *impianto di smaltimento*).
- **Punto di conferimento** (17 concetti). Comprende quei concetti che si riferiscono a luoghi e contenitori dove si conferiscono i rifiuti. Tra di essi troviamo il *bidone carrellato*, il *cassone scarrabile* e il *cassonetto stradale*.
- **Gestione ambientale** (12 concetti). Contiene il concetto di *gestione del ciclo integrato dei rifiuti* definito come "il complesso delle attività volte ad ottimizzare la gestione dei rifiuti urbani". Tale concetto si divide a sua volta in concetti più specifici come quello di *gestione tariffe e rapporto con gli utenti* e *gestione del servizio di raccolta*.
- **Conferimento** (5 concetti). Include i concetti che si riferiscono alla diverse modalità tramite cui i rifiuti vengono consegnati al circuito di raccolta. Tra di esse troviamo il *conferimento diretto*, cioè presso il centro di raccolta, e il *conferimento fuori porta* che corrisponde al ritiro davanti alla propria abitazione.
- **Utenza** (5 concetti). Include i concetti relativi ai diversi tipi di utenze della tassa rifiuti: l'*utenza domestica*, l'*utenza non domestica* e le ulteriori suddivisioni operate a livello locale.
- **Gestore** (4 concetti). Si riferisce al soggetto che gestisce il servizio di igiene urbana. Inoltre, include i concetti relativi ai gestori dei singoli servizi, come il *gestore della raccolta e trasporto* e il *gestore dello spazzamento e lavaggio*.
- **Figura professionale** (4 concetti). Comprende i concetti relativi alle diverse figure professionali che si occupano di gestione dei rifiuti. Tra di esse troviamo l'*operatore ecologico* e il *responsabile del centro di raccolta*.

6.5 Dimensione semantico-comunicativa

Dal punto di vista semantico-comunicativo, ItaIst-WasteLexicon presenta una notevole variabilità terminologica: ogni concetto ha in media 2,4 denominazioni. In particolare, i concetti con il maggior numero di varianti sono quelli relativi alle frazioni merceologiche oggetto di raccolta differenziata. Ad esempio, il concetto di *rifiuto biodegradabile da giardini e parchi* presenta ben 33 denominazioni diverse, tra cui *rifiuto verde*, *sfalci e potature*, *residuo vegetale*, ecc., il *rifiuto urbano residuo* presenta

18 denominazioni: *indifferenziato*, *rifiuto secco residuo*, *frazione non differenziabile*, ecc., la *carta e cartone* ha 19 denominazioni e il *rifiuto organico* ne ha 14.

Per determinare empiricamente il grado di specialismo delle varianti terminologiche (Serianni, 2007; Vellutino, 2018), abbiamo sviluppato una metrica che misura l'associazione tra una determinata variante terminologica e un contesto comunicativo. Se, in questo calcolo consideriamo come contesti comunicativi la comunicazione tra esperti e la comunicazione tra esperti e non esperti, otteniamo una stima del grado di specialismo di un termine, chiameremo questa metrica *rilevanza terminologica*. Infine, a seguito delle analisi esposte in questo paragrafo, abbiamo assegnato ai termini delle note che ne sottolineano l'appropriatezza (o meno) a seconda dei contesti comunicativi "atti amministrativi" e "testi informativi" assegnando, per ogni categoria, una delle seguenti etichette:

- **preferred:** il termine è altamente consigliato in quel contesto comunicativo;
- **admitted:** il termine è ammesso in quel contesto comunicativo;
- **deprecated:** il termine è obsoleto o altamente sconsigliato in quel contesto comunicativo.

6.5.1 Rilevanza terminologica

La rilevanza terminologica confronta la frequenza osservata di una data variante terminologica in un dato contesto comunicativo con la sua frequenza attesa, sotto l'ipotesi che la preferenza per una determinata variante terminologica, tra le possibili, sia indipendente dal contesto comunicativo considerato.

Nello specifico, la formula per calcolare la rilevanza è la seguente:

$$rilevanza(v,t) = F_{v,t} - E_{v,t} \tag{6.1}$$

dove $F_{v,t}$ è la frequenza osservata della variante terminologica v nel contesto comunicativo t mentre $E_{v,t}$ è la frequenza attesa della variante terminologica v nel contesto comunicativo t, sotto l'ipotesi che la probabilità di una data variante terminologica sia indipendente dal contesto comunicativo considerato. $E_{v,t}$ è calcolata con la formula seguente

$$E_{v,t} = \sum_{c \in C} F_{c,t} \cdot \frac{F_v}{\sum_{d \in C} F_d} \tag{6.2}$$

dove $F_{c,t}$ è la frequenza della variante terminologica c nel contesto comunicativo t; F_v e F_d sono rispettivamente la frequenza della variante terminologica v e della variante terminologica d in tutti i contesti comunicativi considerati. Infine, C è l'insieme delle varianti terminologiche che esprimono un determinato concetto.

In questo esperimento, abbiamo calcolato la rilevanza dei termini di ItaIst-WasteLexicon considerando come contesti comunicativi i due sotto-corpora di ItaIst-TermRifiuti: atti amministrativi e testi informativi. Date le ridotte dimensioni del corpus, è stato possibile effettuare l'analisi solo per i termini più frequenti (con almeno 6 occorrenze). Per questo motivo, abbiamo effettuato un'ulteriore analisi sul corpus ItaIst-DdAC_GRU considerando le varianti terminologiche con almeno 10 occorrenze.

Dato che le macrotipologie testuali considerate sono solo due, è sufficiente analizzare la rilevanza relativa ai soli atti amministrativi per avere una stima del grado di specialismo delle varianti terminologiche. Un valore maggiore di zero indica un'associazione positiva con gli atti amministrativi e,

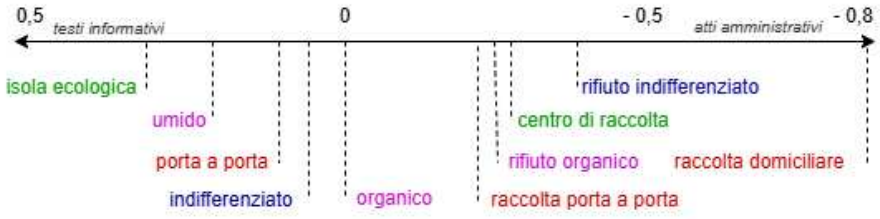


Figura 6.2: Gradi di specialismo in ItaIst-Termrifiuti.

di conseguenza, un grado di specialismo elevato. Al contrario, un valore inferiore a zero indica un'associazione negativa con gli atti amministrativi e positiva con i testi informativi. Di conseguenza, tale valore suggerisce un grado di specialismo basso.

6.5.2 Risultati

L'analisi del grado di specialismo attraverso il calcolo della rilevanza ha evidenziato come alcuni termini siano particolarmente associati agli atti amministrativi (e quindi abbiano un grado di specialismo elevato), o al contrario siano più associati ai testi informativi (e quindi abbiano un basso grado di specialismo).

Tuttavia, tali risultati devono essere interpretati con cautela, alla luce delle dimensioni contenute del campione e dell'influenza del singolo testo sul conteggio delle occorrenze di una data variante. Ad esempio, il basso grado di specialismo associato agli acronimi *CCR* (*Centro Comunale di Raccolta*) e *PAP* (*Porta A Porta*) è dovuto a pochi testi informativi.

Il grado di specialismo delle varianti terminologiche ritenute più interessanti, ovvero quelle che esprimono i concetti di *centro di raccolta*, *raccolta differenziata porta a porta*, *rifiuto urbano residuo* e *rifiuto organico*, è riportato in Figura 6.2.

Emerge che, tra le varianti terminologiche a basso grado di specialismo, sono presenti numerosi aggettivi sostantivati, ottenuti con l'omissione del nome. Invece, nei termini ad alto grado di specialismo tale nome è presente. Ad esempio: *il rifiuto organico* → *l'organico*; *il rifiuto indifferenziato* → *l'indifferenziato*; *la raccolta porta a porta* → *il porta a porta*. Inoltre, nei testi informativi troviamo spesso il termine *isola ecologica* in luogo del termine ufficiale *centro di raccolta* mentre a volte, negli atti amministrativi, si trova il termine *raccolta domiciliare* invece di *raccolta porta a porta*.

L'analisi effettuata su ItaIst-DdAC_GRU, i cui risultati sono riportati in Appendice B - Gradi di specialismo in ItaIst-DdAC_GRU, conferma solo in parte i risultati appena esposti. Infatti, se analizziamo le varianti terminologiche relative ai concetti di *cassone scarrabile* e *bidone carrellato*, possiamo osservare come, in questi due casi, sono gli aggettivi sostantivati *scarrabile* e *carrellato* ad avere un grado di specialismo più alto (rispettivamente 1,5 e 2,9) mentre i termini poliremotici *cassone scarrabile* e *bidone carrellato* hanno un grado di specialismo inferiore (rispettivamente -2,4 e 1,4).

Alla luce di questi risultati, possiamo supporre che gli aggettivi sostantivati siano varianti terminologiche di basso specialismo solo quando tali aggettivi sono entrati stabilmente nell'uso comune. Al contrario, quando gli aggettivi sostantivati non sono di uso comune, tali aggettivi sono varianti di alto specialismo mentre le forme estese *nome + aggettivo* presentano un basso grado di specialismo.

```
SELECT DISTINCT ?text
WHERE {
    ?concept rdfs:label "centro di raccolta"@it.
    ?entry ontolex:evokes ?concept .
    ?entry ontolex:canonicalForm/ontolex:writtenRep ?text .
    }
```

Figura 6.3: Ricerca dei termini che esprimono un dato concetto.

```
SELECT ?concept (GROUP_CONCAT (DISTINCT ?entry; separator=', ') AS ?entries)
WHERE {
    ?concept skos:broader+ wastelexicon:concept_impianto_di_trattamento ;
    ontolex:isEvokedBy/ontolex:canonicalForm/ontolex:writtenRep ?entry .
}
GROUP BY ?concept
```

Figura 6.4: Ricerca ricorsiva dei concetti subordinati.

6.6 Formati di rappresentazione e modalità di interrogazione

Per aderire il più possibile ai principi FAIR (Wilkinson et al., 2016), la base di dati terminologica ItaIst-WasteLexicon è distribuita su Zenodo[5] (Cirillo et al., 2025) con licenza CC-BY 4.0 in formato RDF e descritta tramite il paradigma *Linguistic Linked Open Data* (LLOD) (Cimiano et al., 2020), ovvero tramite le ontologie *Simple Knowledge Organization System* (SKOS) (W3C, 2009) e *OntoLex-Lemon* (Cimiano et al., 2016; McCrae et al., 2017).

Inoltre, abbiamo messo a disposizione un *endpoint* SPARQL[6] grazie al quale è possibile interrogare la base dati terminologica. Il punto di forza dell'*endpoint* SPARQL risiede nel fatto che permette di eseguire numerosi tipi di *query*, anche complessi e articolati, e di ottenere l'output in diversi formati, sia *human-readable* che *machine-readable*.

Di seguito sono riportate alcune query che esemplificano le potenzialità di interrogazione di ItaIst-WasteLexicon. La prima *query*, illustrata in Figura 6.3 restituisce tutti i termini che esprimono un dato concetto (che nell'esempio è *centro di raccolta*). Invece, la query in Figura 6.4 restituisce tutti i concetti subordinati a un dato concetto (che nell'esempio è *impianto di trattamento*).

Infine, la connessione tra ItaIst-WasteLexicon e ItaIst-TermRifiuti rende possibile eseguire *query* combinate tra le due risorse. Ad esempio, con la *query* in Figura 6.5 è possibile ricercare tutte le frasi in cui è presente un termine che esprime un dato concetto, che nell'esempio è *centro di raccolta*.

[5] https://doi.org/10.5281/zenodo.15173712
[6] http://193.205.161.60:8890/sparql/ (Ultima rilevazione: 10/04/2025)

```
SELECT DISTINCT ?term ?sentence
WHERE {
    ?concept rdfs:label "centro di raccolta"@it;
    ontolex:lexicalizedSense/^itsrdf:termInfoRef ?annotation .
    ?annotation powla:string ?term ;
    powla:hasChild{2}/^nif:word/nif:anchorOf ?sentence .
    }
```

Figura 6.5: Ricerca di un concetto in ItaIst-TermRifiuti

Capitolo 7

Conclusioni

L'uso corretto della terminologia istituzionale è fondamentale per la buona riuscita della comunicazione tra le istituzioni e i loro pubblici di riferimento. In particolare, la terminologia è un tema chiave nel contesto della gestione dei rifiuti urbani, dove la comunicazione gioca un ruolo di rilievo nel raggiungere gli obiettivi ambientali fissati dalla normativa. Riteniamo quindi che lo studio della terminologia istituzionale della gestione dei rifiuti urbani possa contribuire a migliorare la qualità dell'italiano istituzionale. Tuttavia, per essere efficace, tale studio deve necessariamente fondarsi su solide basi empiriche, sotto forma di risorse linguistico-terminologiche che siano facilmente accessibili ed interoperabili con altre risorse, sia linguistiche che di altra natura.

Per questo motivo, abbiamo sviluppato quattro risorse linguistico-terminologiche: i corpora specialistici ItaIst-DdAC_GRU e ItaIst-TermRifiuti, la tecnica di estrazione automatica di termini *Domain Concept Relatedness* e la base di dati terminologica ItaIst-WasteLexicon. Tutte le risorse sono state realizzate ponendo particolare attenzione al rispetto dei paradigmi FAIR e *Linguistic Linked Open Data*.

Le risorse sviluppate rappresentano materiale di studio non solo per la terminologia istituzionale della gestione dei rifiuti urbani, ma anche, più in generale, per la terminologia computazionale, in particolare per l'estrazione automatica di termini, e per lo studio dell'italiano istituzionale. Inoltre, il presente lavoro ha fornito un contribuito allo sviluppo di una metodologia con la quale realizzare risorse linguistico-terminologiche FAIR che possono trovare ulteriori impieghi, che vadano anche al di là del loro scopo originario.

Infine, l'analisi della base di dati terminologica ItaIst-WasteLexicon e dei corpora ItaIst-DdAC_GRU e ItaIst-TermRifiuti ha confermato i tratti rilevati nei precedenti studi sul profilo dell'italiano istituzionale e evidenziato le seguenti caratteristiche della terminologia istituzionale della gestione dei rifiuti urbani:

- L'alta variabilità terminologica. Specialmente associata ai nomi delle frazioni merceologiche oggetto di raccolta differenziata, che solo in alcuni casi è dovuta a reali esigenze comunicative. Ad esempio, il rifiuto urbano residuo presenta ben 18 denominazioni all'interno di ItaIst-WasteLexicon (es. *indifferenziato*, *rifiuto secco residuo*, *frazione non differenziabile*, ecc.)

- La presenza di aggettivi sostantivati. Tali termini, ottenuti con l'omissione del nome testa (es. *organico*, *differenziata*, *porta a porta*), sono termini a basso grado di specialismo che appaiono di frequente nei testi informativi. Al contrario, nei testi con fuzione prescrittiva, come gli atti

amministrativi, il nome testa è presente (es. *rifiuto organico*, *raccolta differenziata*, *raccolta porta a porta*). Alcuni di questi aggettivi sostantivati, che non sono di uso comune, pongono problemi di comprensibilità. Ad esempio, *scarrabile* da *cassone scarrabile* e *carrellato* da *bidone carrellato*.

- La presenza di nomi deverbali e nomi a suffissazione zero. Dall'analisi della morfologia dei termini monorematici è emerso che quasi tutti i nomi e gli aggettivi derivati sono deverbali e che molti di essi sono derivati per suffisazione zero, come, ad esempio, *deposito* da *depositare*, *riciclo* da *riciclare* e *recupero* da *recuperare*.

- La presenza di perifrasi eccessivamente complesse. Grazie all'analisi dei termini polirematci abbiamo osservato la presenza di complesse perifrasi che diminuiscono la leggibilità dei testi. Come ad esempio la perifrasi *rifiuto proveniente da luogo adibito ad uso di civile abitazione* utilizzata in sostituzione del termine *rifiuto domestico*, decisamente più semplice.

La limitata copertura geografica dei corpora non ha permesso di catturare tutte le varianti terminologiche né di individuare varianti diatopiche che probabilmente sono rilevanti per la terminologia istituzionale della gestione dei rifiuti urbani.

In futuro, prevediamo quindi di mantenere aggiornate le risorse inserendo nuovi dati e ampliando la copertura geografica, rendendo così possibile un'analisi diacronica e diatopica. Inoltre, intendiamo analizzare, con la stessa metodologia, altri domini terminologici dei linguaggi istituzionali e compararli con quello della gestione dei rifiuti urbani, in modo da identificarne similarità e differenze.

Bibliografia

Ahmad, K., Gillam, L., Tostevin, L., et al. University of surrey participation in trec8: Weirdness indexing for logical document extrapolation and retrieval (wilder). In: In *TREC*. 1999, 1–8.

Arcan, M., Turchi, M., Topelli, S., & Buitelaar, P. Enhancing statistical machine translation with bilingual terminology in a CAT environment (Y. Al-Onaizan & M. Simard, Cur.). In: *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: MT Researchers Track* (Y. Al-Onaizan & M. Simard, Cur.). A cura di Al-Onaizan, Y., & Simard, M. Association for Machine Translation in the Americas, 2014, ottobre, 54–68.

Astrakhantsev, N. (2018). ATR4S: toolkit with state-of-the-art automatic terms recognition methods in Scala. *Language Resources and Evaluation*, *52*(3), 853–872.

Astrakhantsev, N., Fedorenko, D., & Turdakov, D. (2015). Methods for automatic term recognition in domain-specific text collections: A survey. *Programming and Computer Software*, *41*(6), 336–349.

Auger, P., & L'Homme, M.-C. La terminologie selon une approache textuelle: Une représentation plus adéquate du lexique dans les langues de spécialité. In: In *ALFA: Actes de Langue Française et de Linguistique / Symposium on French Language and Linguistics*. 1994/1995, 17–21.

Banerjee, S., Chakravarthi, B. R., & McCrae, J. P. Large Language Models for Few-Shot Automatic Term Extraction (A. Rapp, L. Di Caro, F. Meziane & V. Sugumaran, Cur.). In: *Natural Language Processing and Information Systems* (A. Rapp, L. Di Caro, F. Meziane & V. Sugumaran, Cur.). A cura di Rapp, A., Di Caro, L., Meziane, F., & Sugumaran, V. Cham: Springer Nature Switzerland, 2024, 137–150. ISBN: 978-3-031-70239-6.

Basili, R., Moschitti, A., Teresa, P. M., & Zanzotto, F. M. A contrastive approach to term extraction. In: 2001. https://www.researchgate.net/publication/283854408

Bellandi, A. (2021). LexO: an open-source system for managing OntoLex-Lemon resources. *Language Resources and Evaluation*, *55*(4), 1093–1126. https://doi.org/10.1007/s10579-021-09546-4

Berruto, G. (1987). *Sociolinguistica dell'italiano contemporaneo* [2ª ed. 2012]. Carocci.

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing*, *8*, 243–257. https://doi.org/10.1093/llc/8.4.243

Biber, D., & Reppen, R. (2015, giugno). The Cambridge Handbook of English Corpus Linguistics. In D. Biber & R. Reppen (Cur.). Cambridge University Press. https://doi.org/10.1017/CBO9781139764377

Bonin, F., Dell'Orletta, F., Montemagni, S., & Venturi, G. A Contrastive Approach to Multi-word Extraction from Domain-specific Corpora. In: In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: Euro-

pean Language Resources Association (ELRA), 2010, maggio. http://www.lrec-conf.org/proceedings/lrec2010/pdf/553_Paper.pdf

Boulanger, J.-C. (1995). Présentation : images et parcours de la socioterminologie. *Meta*, *40*, 194–205. https://doi.org/10.7202/002117ar

Bourigault, D., & Slodzian, M. (1999). Pour une terminologie textuelle. *Terminologies nouvelles*, *19*(1999), 29–32.

Bowker, L., & Pearson, J. (2002). *Working with specialized language: A practical guide to using corpora*. Routledge.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. Language Models are Few-Shot Learners (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan & H. Lin, Cur.). In: *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan & H. Lin, Cur.). A cura di Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., & Lin, H. *33*. Curran Associates, Inc., 2020, 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Cabré Castellví, M. T. (1998). Do we need an autonomous theory of terms? *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *5*, 4–19. https://doi.org/10.1075/term.5.1.03cab

Cabré Castellví, M. T. (1999). *Terminology : Theory, Methods, and Applications* (J. C. Sager, Cur.). John Benjamins. https://doi.org/https://doi.org/10.1075/tlrp.1

Cabré Castellví, M. T. (2003). Theories of terminology. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *9*, 163–199. https://doi.org/10.1075/term.9.2.03cab

Campo, (2013). The reception of Eugen Wüster's work and the development of terminology.

Chiarcos, C. Ontologies of Linguistic Annotation: Survey and perspectives (N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis, Cur.). In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis, Cur.). A cura di Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., & Piperidis, S. Istanbul, Turkey: European Language Resources Association (ELRA), 2012, maggio, 303–310. https://aclanthology.org/L12-1545/

Chiarcos, C. POWLA: Modeling Linguistic Corpora in OWL/DL (E. Simperl, P. Cimiano, A. Polleres, O. Corcho & V. Presutti, Cur.). In: *The Semantic Web: Research and Applications* (E. Simperl, P. Cimiano, A. Polleres, O. Corcho & V. Presutti, Cur.). A cura di Simperl, E., Cimiano, P., Polleres, A., Corcho, O., & Presutti, V. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, 225–239. ISBN: 978-3-642-30284-8.

Church, K., & Gale, W. Inverse document frequency (idf): A measure of deviations from poisson. In: In *Natural language processing using very large corpora*. Springer, 1999, pp. 283–295.

Cimiano, P., Chiarcos, C., McCrae, J. P., & Gracia, J. (2020). *Linguistic Linked Data Representation, Generation and Applications* (1ª ed.). Springer Cham. https://doi.org/10.1007/978-3-030-30225-2

Cimiano, P., McCrae, J. P., & Buitelaar, P. (2016). *Lexicon Model for Ontologies: Community Report* (Community Report). World Wide Web Consortium. https://www.w3.org/2016/05/ontolex/

Cimmino, D., Maisto, A., Falco, M., Pelosi, S., Catone, M., & Paone, A. (2022). *Cultural Heritage Corpus, uno sguardo testuale sul turismo emozionale* [Ultima rilevazione: 10/04/2025]. https://culturalheritageunisa.wordpress.com/

Cirillo, N. (2024). Isolating Terminology Layers in Complex Linguistic Environments: A Study about Waste Management. *Languages*, *9*, 68. https://doi.org/10.3390/languages9030068

Cirillo, N., & Vellutino, D. Term Variation in Institutional Languages: Degrees of Specialization in Municipal Waste Management Terminology. In: In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context@ LREC-COLING 2024*. 2024, 134–140.

Cirillo, N., Vellutino, D., Nicoletti, D., Sabarese, E., & Rubino, B. (2025, marzo). ItaIst_GRU. https://doi.org/10.5281/zenodo.15173712

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. Unsupervised Cross-lingual Representation Learning at Scale (D. Jurafsky, J. Chai, N. Schluter & J. Tetreault, Cur.). In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (D. Jurafsky, J. Chai, N. Schluter & J. Tetreault, Cur.). A cura di Jurafsky, D., Chai, J., Schluter, N., & Tetreault, J. Online: Association for Computational Linguistics, 2020, luglio, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

Cortelazzo, M. (2021). *Il linguaggio amministrativo. Principi e pratiche di modernizzazione*. Carocci.

De Mauro, T., & Chiari, I. (2016). *Il Nuovo vocabolario di base della lingua italiana*. https://www.internazionale.it/opinione/tullio-de-mauro/2016/12/23/il-nuovo-vocabolario-di-base-della-lingua-italiana

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (J. Burstein, C. Doran & T. Solorio, Cur.). In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran & T. Solorio, Cur.). A cura di Burstein, J., Doran, C., & Solorio, T. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, giugno, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Diki-Kidiri, M. (2022). Chapter 9. Cultural Terminology: An introduction to theory and method. In P. Faber & M.-C. L'Homme (Cur.), *Theoretical Perspectives on Terminology: Explaining terms, concepts and specialized knowledge* (pp. 197–216). John Benjamins Publishing Company. https://doi.org/doi:10.1075/tlrp.23.09dik

Donati, M., Polidori, L., Vernillo, P., Gagliardi, G., et al. Building a corpus on Eating Disorders from TikTok: challenges and opportunities. In: In *CEUR WORKSHOP PROCEEDINGS. 3596*. Aachen. 2023, 1–6.

Drouin, P., L'Homme, M.-C., & Robichaud, B. Lexical Profiling of Environmental Corpora (N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis & T. Tokunaga, Cur.). In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara,

B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis & T. Tokunaga, Cur.). A cura di Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., & Tokunaga, T. European Language Resources Association (ELRA), 2018, maggio. https: //aclanthology.org/L18-1539

Drouin, P., L'Homme, M.-C., & Robichaud, B. Lexical profiling of environmental corpora. In: In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.

El-Beltagy, S. R., & Rafea, A. KP-Miner: Participation in SemEval-2. In: Association for Computational Linguistics, 2010, luglio, 190–193. https://aclanthology.org/S10-1041

Erjavec, T., Kopp, M., Ogrodniczuk, M., Osenova, P., Agirrezabal, M., Agnoloni, T., Aires, J., Albini, M., Alkorta, J., Antiba-Cartazo, I., Arrieta, E., Barcala, M., Bardanca, D., Barkarson, S., Bartolini, R., Battistoni, R., Bel, N., Bonet Ramos, M. d. M., Calzada Pérez, M., ... Fišer, D. (2024). Multilingual comparable corpora of parliamentary debates ParlaMint 4.1 [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/1912

Evans, D. A., & Lefferts, R. G. (1995). CLARIT - TREC experiments. *Information Processing & Management*, *31*(3), 385–395.

Faber, P., León-Araúz, P., & Reimerink, A. EcoLexicon: New Features and Challenges. In: In *GLOBALEX 2016 Lexicographic Resources for Human Language Technology Workshop Programme*. 2016, 73.

Faber, P., & L'Homme, M.-C. (Cur.). (2022). *Theoretical Perspectives on Terminology*. John Benjamins. https://www.jbe-platform.com/content/books/9789027257789

Fedorenko, D., Astrakhantsev, N., & Turdakov, D. (2014). Automatic Recognition of Domain-Specific Terms: an Experimental Evaluation. *Proceedings of the Institute for System Programming of RAS*, *26*, 55–72. https://doi.org/10.15514/ISPRAS-2014-26(4)-5

Fillmore, C. J. (1985). Problèmes terminologiques des pluies acides : pour une socio-terminologie. *Quaderni di Semantica*, *32*, 222–254.

Fiorelli, M., Stellato, A., Lorenzetti, T., Turbati, A., Schmitz, P., Francesconi, E., Hajlaoui, N., & Batouche, B. (2018). Towards OntoLex–Lemon editing in VocBench 3. *AIDAinformazioni*, *2018*(Numero speciale), 81–102.

Fiorentino, G., & Ganfi, V. (2024). Parametri per semplificare l'italiano istituzionale: revisione della letteratura. *Italiano LinguaDue*, *16*(1), 220–237.

Fioritto, A. (1997). *Manuale di stile. Strumenti per semplificare il linguaggio delle amministrazioni pubbliche*. Il Mulino.

Gambier, Y. (1998). Problèmes terminologiques des pluies acides : pour une socio-terminologie. *Meta*, *32*, 314–320. https://doi.org/10.7202/002791ar

Gaudin, F. (1993). *Socioterminologie des problems semantiques aux pratiques institutionnelles*. Publications de l'Université de Rouen.

Gaudin, F. (2003). *Socioterminologie: Une approche sociolinguistique de la terminologie*. De Boeck Larcier.

Giovannetti, E., Albanesi, D., Bellandi, A., Marchi, S., Papini, M., & Sciolette, F. (2024). Maia: an Open Collaborative Platform for Text Annotation, E-Lexicography, and Lexical Linking. *Umanistica Digitale*, *8*(18), 27–52. https://doi.org/10.6092/issn.2532-8816/19705

Grandi, N., Ballarè, S., Chiusaroli, F., Gallina, F., Pascoli, M., & Pistolesi, E. (2023). *Corpus UniverS-Ita*. https://doi.org/https://doi.org/10.60760/unibo/univers-ita

Gualdo, R., & Telve, S. (2021). *Linguaggi specialistici dell'italiano*. Carocci.

Hazem, A., Bouhandi, M., Boudin, F., & Daille, B. Termeval 2020: Taln-ls2n system for automatic term extraction. In: In *6th International Workshop on Computational Terminology (COMPUTERM 2020)*. 2020.

Hellmann, S., Lehmann, J., Auer, S., & Brümmer, M. Integrating NLP Using Linked Data (H. Alani, L. Kagal, A. Fokoue, P. Groth, C. Biemann, J. X. Parreira, L. Aroyo, N. Noy, C. Welty & K. Janowicz, Cur.). In: *The Semantic Web – ISWC 2013* (H. Alani, L. Kagal, A. Fokoue, P. Groth, C. Biemann, J. X. Parreira, L. Aroyo, N. Noy, C. Welty & K. Janowicz, Cur.). A cura di Alani, H., Kagal, L., Fokoue, A., Groth, P., Biemann, C., Parreira, J. X., Aroyo, L., Noy, N., Welty, C., & Janowicz, K. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, 98–113. ISBN: 978-3-642-41338-4.

ISO 1087:2019. (2019). *ISO 1087:2019 - Terminology work and terminology science*. International Organization for Standardization.

ISO 30042:2019. (2019). *ISO 30042:2019 - Management of terminology resources — TermBase eXchange (TBX)*. International Organization for Standardization.

Justeson, J. S., & Katz, S. M. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural language engineering*, *1*(1), 9–27.

Khodak, M., Saunshi, N., Liang, Y., Ma, T., Stewart, B., & Arora, S. A La Carte Embedding: Cheap but Effective Induction of Semantic Feature Vectors. In: In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, 2018, luglio, 12–22. https://doi.org/10.18653/v1/P18-1002

Kilgarriff, A. Simple maths for keywords. In: In *Proc. Corpus Linguistics. 6*. 2009.

Kim, J. D., Ohta, T., Tateisi, Y., & Tsujii, J. GENIA corpus - A semantically annotated corpus for bio-textmining. In: In *Bioinformatics. 19*. Oxford University Press, 2003. https://doi.org/10.1093/bioinformatics/btg1023

Koehn, P. Europarl: A Parallel Corpus for Statistical Machine Translation. In: In *Proceedings of Machine Translation Summit X: Papers*. Phuket, Thailand, 2005, 79–86. https://aclanthology.org/2005.mtsummit-papers.11/

Kucza, M., Niehues, J., Zenkel, T., Waibel, A., & Stüker, S. Term extraction via neural sequence labeling a comparative evaluation of strategies using recurrent neural networks. In: *2018-September*. International Speech Communication Association, 2018, 2072–2076. https://doi.org/10.21437/Interspeech.2018-2017

Lakoff, G. (1987). *Women, Fire and Dangerous Things*. University of Chicago Press.

Lang, C., Wachowiak, L., Heinisch, B., & Gromann, D. Transforming Term Extraction: Transformer-Based Approaches to Multilingual Term Extraction Across Domains (C. Zong, F. Xia, W. Li & R. Navigli, Cur.). In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (C. Zong, F. Xia, W. Li & R. Navigli, Cur.). A cura di Zong, C., Xia, F., Li, W., & Navigli, R. Online: Association for Computational Linguistics, 2021, agosto, 3607–3620. https://doi.org/10.18653/v1/2021.findings-acl.316

Lenci, A., Montemagni, S., Pirrelli, V., & Venturi, G. Ontology learning from Italian legal texts. In: In *Law, Ontologies and the Semantic Web*. IOS Press, 2009, pp. 75–94.

Leoń-Arauź, P., Martiń, A. S., & Reimerink, A. The EcoLexicon English Corpus as an open corpus in Sketch Engine. In: In *EURALEX Proceedings*. European Association for Lexicography, 2018, 893–901. ISBN: 9789610600978.

L'Homme, M.-C. (2006). The processing of terms in dictionaries: New models and techniques. A state of the art. *Terminology*, *12*, 181–188.

Lubello, S. (2014a). *Il linguaggio burocratico*. Carocci.

Lubello, S. (2014b). *Lezioni d'italiano: riflessioni sulla lingua del nuovo millennio*. Il mulino.

Lubello, S. L'antilingua gode di buona salute: nuove forme, vecchi vizi. In: In *Comunicare cittadinanza nell'era digitale Saggi sul linguaggio burocratico 2.0*. FrancoAngeli, 2018, pp. 31–43.

Lyding, V., Stemle, E., Borghetti, C., Brunello, M., Castagnoli, S., Dell'orletta, F., Dittmann, H., Lenci, A., & Pirrelli, V. The PAISA' Corpus of Italian Web Texts (F. Bildhauer & R. Schäfer, Cur.). In: a cura di Bildhauer, F., & Schäfer, R. Association for Computational Linguistics, 2014, aprile, 36–43.

Manjunath, S. H., & McCrae, J. P. Encoder-Attention-Based Automatic Term Recognition (EA-ATR). In: In *3rd Conference on Language, Data and Knowledge (LDK 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik. 2021.

Martín-Chozas, P., & Declerck, T. Representing Multilingual Terminologies with OntoLex-Lemon (short paper) (G. M. D. Nunzio, G. M. Henrot, M. T. Musacchio & F. Vezzani, Cur.). In: *Proceedings of the 1st International Conference on Multilingual Digital Terminology Today, Padua, Italy, June 16-17, 2022 (hybrid event)* (G. M. D. Nunzio, G. M. Henrot, M. T. Musacchio & F. Vezzani, Cur.). A cura di Nunzio, G. M. D., Henrot, G. M., Musacchio, M. T., & Vezzani, F. *3161*. CEUR Workshop Proceedings. CEUR-WS.org, 2022. https://ceur-ws.org/Vol-3161/short1.pdf

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., & Cimiano, P. The OntoLex-Lemon Model: Development and Applications. In: In *Proceedings of eLex 2017*. 2017, 587–597.

Meijer, K., Frasincar, F., & Hogenboom, F. (2014). A Semantic Approach for Extracting Domain Taxonomies from Text. *Decision Support Systems*, *62*. https://doi.org/10.1016/j.dss.2014.03.006

Meyers, A. L., He, Y., Glass, Z., Ortega, J., Liao, S., Grieve-Smith, A., Grishman, R., & Babko-Malaya, O. (2018). The termolator: terminology recognition based on chunking, statistical and search-based scores. *Frontiers in Research Metrics and Analytics*, 19.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *1st International Conference on Learning Representations, {ICLR} 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Miles, A., & Bechhofer, S. (2009). *SKOS Simple Knowledge Organization System eXtension for Labels (SKOS-XL) Namespace Document - HTML Variant* (rapp. tecn.). World Wide Web Consortium. http://www.w3.org/TR/2009/REC-skos-reference-20090818/skos-xl.html

Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., & Liang, X. (2018). doccano: Text Annotation Tool for Human [Software available from https://github.com/doccano/doccano]. https://github.com/doccano/doccano

Navigli, R., & Velardi, P. (2004). Learning domain ontologies from document warehouses and dedicated web sites. *Computational Linguistics*, *30*(2), 151–179.

Oliver, A., & Vàzquez, M. Termeval 2020: Using tsr filtering method to improve automatic term extraction. In: In *Proceedings of the 6th International Workshop on Computational Terminology*. 2020, 106–113.

Pais, V., & Ion, R. TermEval 2020: RACAI's automatic term extraction system. English. In: In *Proceedings of the 6th International Workshop on Computational Terminology*. Marseille, France: European Language Resources Association, 2020, maggio, 101–105. ISBN: 979-10-95546-57-3. https://aclanthology.org/2020.computerm-1.14

Park, Y., Byrd, R. J., & Boguraev, B. Automatic Glossary Extraction: Beyond Terminology Identification. In: In *COLING. 10*. 2002, 1072228–1072370.

Patry, A., & Langlais, P. Corpus-based terminology extraction. In: In *Terminology and Content Development–Proceedings of 7th International Conference On Terminology and Knowledge Engineering, Litera, Copenhagen*. 2005.

Piccini, S., Vezzani, F., & Bellandi, A. (2023). TBX and 'Lemon': What perspectives in terminology? *Digital Scholarship in the Humanities*, *38*(Supplement$_1$), i61–i72. https://doi.org/10.1093/llc/fqad025

Piemontese, M. E. Scrittura e leggibilità: 'due parole' (M. Cortelazzo, Cur.). In: *Scrivere nella scuola dell'obbligo, Quaderni del Giscel* (M. Cortelazzo, Cur.). A cura di Cortelazzo, M. La Nuova Italia, 1999, pp. 151–167.

Qasemizadeh, B., & Schumann, A.-K. The ACL RD-TEC 2.0: A Language Resource for Evaluating Term Extraction and Entity Recognition Methods (N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis, Cur.). In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis, Cur.). A cura di Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., & Piperidis, S. European Language Resources Association (ELRA), 2016, maggio, 1862–1868. ISBN: 0.635413337.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In: In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 2020. https://nlp.stanford.edu/pubs/qi2020stanza.pdf

Restivo, M. L. (2022). L'Italiano scritto degli studenti universitari: prime osservazioni sul corpus univers-ita. *Italiano LinguaDue*, *14*(1), 797–818.

Rigouts Terryn, A., Hoste, V., & Lefever, E. (2020). In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora. *Language Resources and Evaluation*, *54*, 385–418. https://doi.org/10.1007/s10579-019-09453-9

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic keyword extraction from individual documents. *Text mining: applications and theory*, *1*, 1–20.

Scansani, R., Bentivogli, L., Bernardini, S., & Ferraresi, A. MAGMATic: A Multi-domain Academic Gold Standard with Manual Annotation of Terminology for Machine Translation Evaluation (M. Forcada, A. Way, B. Haddow & R. Sennrich, Cur.). In: *Proceedings of Machine Translation Summit XVII: Research Track* (M. Forcada, A. Way, B. Haddow & R. Sennrich, Cur.). A cura di Forcada, M., Way, A., Haddow, B., & Sennrich, R. European Association for Machine Translation, 2019, agosto, 78–86.

Serianni, L. (2007). Terminologia medica: qualche considerazione tra italiano, francese, spagnolo. *Terminologie specialistiche e tipologie testuali*, 7–9.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford University Press.

Speranza, G., di Buono, M. P., & Monti, J. (2023). Towards terminological resources tailored to the users' needs: Terminology extraction based on appositive constructions. *Digital Scholarship in the Humanities*, *38*, i83–i92. https://doi.org/10.1093/llc/fqad024

Stellato, A., Fiorelli, M., Turbati, A., Lorenzetti, T., van Gemert, W., Dechandon, D., Laaboudi-Spoiden, C., Gerencsér, A., Waniart, A., Costetchi, E., & Keizer, J. (2020). VocBench 3: A collaborative Semantic Web editor for ontologies, thesauri and lexicons. *Semantic Web*, *11*(5), 855–881. https://doi.org/10.3233/SW-200370

Stellato, A., Rajbhandari, S., Turbati, A., Fiorelli, M., Caracciolo, C., Lorenzetti, T., Keizer, J., & Pazienza, M. T. VocBench: A Web Application for Collaborative Development of Multilingual Thesauri (F. Gandon, M. Sabou, H. Sack, C. d'Amato, P. Cudré-Mauroux & A. Zimmermann, Cur.). In: *The Semantic Web. Latest Advances and New Domains* (F. Gandon, M. Sabou, H. Sack, C. d'Amato, P. Cudré-Mauroux & A. Zimmermann, Cur.). A cura di Gandon, F., Sabou, M., Sack, H., d'Amato, C., Cudré-Mauroux, P., & Zimmermann, A. Cham: Springer International Publishing, 2015, 38–53. ISBN: 978-3-319-18818-8.

Subirats-Coll, I., Kolshus, K., Turbati, A., Stellato, A., Mietzsch, E., Martini, D., & Zeng, M. (2022). AGROVOC: The linked data concept hub for food and agriculture. *Computers and Electronics in Agriculture*, *196*, 105965. https://doi.org/https://doi.org/10.1016/j.compag.2020.105965

TEI. (2025). *Guidelines for Electronic Text Encoding and Interchange. Version 4.9.0* (rapp. tecn.). TEI Consortium. http://www.tei-c.org/Guidelines/P5/

Temmerman, R. (2000). *Towards New Ways of Terminology Description: the Sociocognitive-approach*. John Benjamins Publishing Company. https://doi.org/https://doi.org/10.1075/tlrp.3

Tran, H. T. H., González-Gallardo, C.-E., Delaunay, J., Doucet, A., & Pollak, S. Is Prompting What Term Extraction Needs? (E. Nöth, A. Horák & P. Sojka, Cur.). In: *Text, Speech, and Dialogue* (E. Nöth, A. Horák & P. Sojka, Cur.). A cura di Nöth, E., Horák, A., & Sojka, P. Cham: Springer Nature Switzerland, 2024, 17–29. ISBN: 978-3-031-70563-2.

Vellutino, D. Esercizi di stile per il diritto di accesso civico. In: In *La lingua variabile nei testi letterari, artistici e funzionali contemporanei. Analisi, interpretazione, traduzione. Atti del XIII Congresso della Società Internazionale di Linguistica e Filologia Italiana*. Centro studi filologici e linguistici siciliani. 2014, 1–16.

Vellutino, D. Risorse linguistiche e Open Data per la comunicazione pubblica della gestione dei rifiuti urbani (D. Vellutino & M. T. Zanola, Cur.). In: *Comunicare in Europa. Lessici istituzionali e terminologie specialistiche* (D. Vellutino & M. T. Zanola, Cur.). A cura di Vellutino, D., & Zanola, M. T. Milano: EDUCatt - Ente per il Diritto allo Studio Universitario dell'Università Cattolica, 2015, pp. 217–245. ISBN: 9788883119163.

Vellutino, D. (2018). *L'italiano istituzionale per la comunicazione pubblica*. il Mulino.

Vellutino, D. (2022). Neologismi istituzionali per la ripresa e resilienza dell'Italia: Analisi tipologico-strutturale dei nuovi termini del PNRR. *AIDAinformazioni*, (3-4), 113–124.

Vellutino, D., & Cirillo, N. (2024). Corpus «itaist»: Note per lo sviluppo di una risorsa linguistica per lo studio dell'italiano istituzionale per il diritto di accesso civico. *Italiano LinguaDue*, *16*(1), 238–250.

Vellutino, D., & Formisano, V. (2016). "MWM- Lexicon". Studio per una risorsa terminologica bilingue sulla "Gestione dei Rifiuti Urbani". *AIDA INFORMAZIONI*, *1*(2), 167–178. https://doi.org/10.4399/978885489402016

Vellutino, D., Marano, F., & Elia, A. L'italiano istituzionale e le sue varietà d'uso pubblico. Aspetti lessicali nei tipi di testo d'informazione e comunicazione delle pubbliche amministrazioni. In: In *Atti XI Congresso SILFI*. 2012.

Vellutino, D., Maslias, R., & Rossi, F. (2016). Verso l'interoperabilità semantica di IATE. Studio preliminare per il dominio "Gestione dei rifiuti urbani". *Terminologie specialistiche e diffusione dei saperi*, 1–240.

Verborgh, R., Röder, M., Usbeck, R., & Ngonga Ngomo, A.-C. (2018). GERBIL – Benchmarking Named Entity Recognition and Linking consistently. *Semantic Web*, *9*(5), 605–625. https://doi.org/10.3233/SW-170286

Vezzani, F. (2021). La ressource FAIRterm : entre pratique pédagogique et professionnalisation en traduction spécialisée The FAIRterm resource: between pedagogical practice and professionalization in specialized translation. *Synergies Italie*, *17*, 51–64.

Vezzani, F. (2022). *Terminologie numérique : conception, représentation et gestion*. Peter Lang Verlag. https://doi.org/10.3726/b19407

W3C. (2009). *SKOS Simple Knowledge Organization System Reference* (rapp. tecn.). World Wide Web Consortium. https://www.w3.org/TR/skos-reference/

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*(1), 1–9.

Wüster, E. (1979). *Einführung in die allgemeine Terminologielehre und terminologische Lexicographic*. Springer.

Zanola, M. T. (2018). *Che cos'è la terminologia*. Carocci.

Zhang, Z., Gao, J., & Ciravegna, F. JATE2. 0: Java Automatic Term Extraction with Apache Solr. In: In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA). 2016.

Appendice A - Esempio di scheda terminologica

isola ecologica

proprietà morfosintattiche

parte del discorso: nome
composizione: nome + aggettivo

forme flesse

isola ecologica	*fem. sing.*
isole ecologiche	*fem. plur*

proprietà concettuali

dominio: gestione dei rifiuti urbani

definizione:
Area presidiata ed allestita, senza nuovi o maggiori oneri a carico della finanza pubblica, per l'attività di raccolta mediante raggruppamento differenziato dei rifiuti urbani per frazioni omogenee conferiti dai detentori per il trasporto agli impianti di recupero e trattamento. La disciplina dei centri di raccolta è data con decreto del Ministro dell'ambiente e della tutela del territorio e del mare, sentita la Conferenza unificata, di cui al decreto legislativo 28 agosto 1997, n. 281.
fonte: DECRETO LEGISLATIVO 3 aprile 2006, n. 152

concetti più generali: impianto di stoccaggio; punto di raccolta;
concetti più specifici: centro di raccolta mobile; centro di raccolta intercomunale

proprietà semantico-comunicative

uso negli atti amministrativi: ammesso
uso nei testi informativi: preferito

varianti terminologiche

relazione	variante	uso	
		atti amministrativi	testi informativi
sinonimo	CCR	ammesso	ammesso
	centro comunale di raccolta	ammesso	ammesso
	centro di raccolta	preferito	ammesso
	centro di raccolta comunale	ammesso	ammesso
	centro di raccolta comunale dei rifiuti urbani	ammesso	ammesso
	centro di raccolta dei rifiuti urbani	ammesso	ammesso
	centro di raccolta dei rifiuti urbani raccolti in modo differenziato	deprecato	deprecato
	centro di raccolta differenziata	ammesso	ammesso
	centro di raccolta rifiuti differenziati	deprecato	deprecato
	centro di raccolta rifiuti urbani	ammesso	ammesso
	ecocentro	ammesso	ammesso
	isola	deprecato	ammesso
iperonimo	centro di stoccaggio	ammesso	ammesso
	impianto di stoccaggio	preferito	preferito
	eco-punto	deprecato	ammesso
	punto di conferimento	ammesso	ammesso
	punto di raccolta	preferito	preferito
	punto raccolta	deprecato	deprecato
iponimo	centro di raccolta intercomunale	ammesso	ammesso
	centro di raccolta rifiuti differenziati intercomunale	deprecato	deprecato
	isola ecologica itinerante	ammesso	preferito

Figura 1: Esempio di scheda terminologica del termine *isola ecologica*

Appendice B - Gradi di specialismo in ItaIst-DdAC_GRU

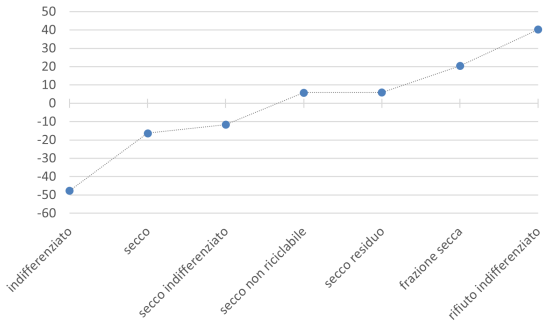


Figura 2: Gradi di specialismo delle varianti terminologiche che esprimono il concetto di "rifiuto urbano residuo"

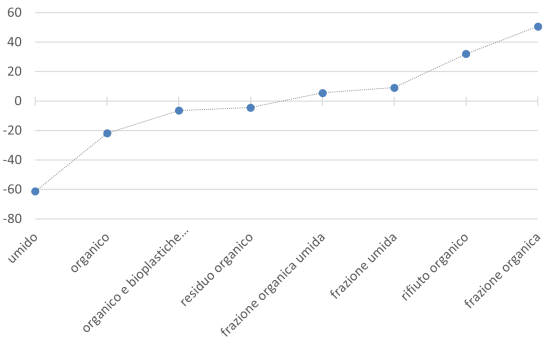


Figura 3: Gradi di specialismo delle varianti terminologiche che esprimono il concetto di "rifiuto organico"

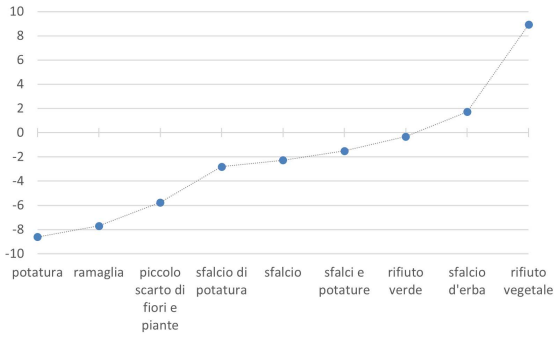


Figura 4: Gradi di specialismo delle varianti terminologiche che esprimono il concetto di "rifiuto biodegradabile di giardini e parchi"

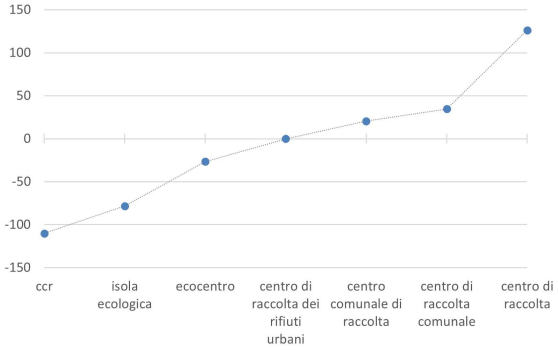


Figura 5: Gradi di specialismo delle varianti terminologiche che esprimono il concetto di "centro di raccolta"

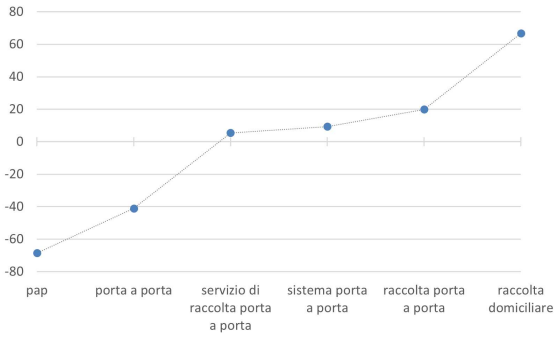


Figura 6: Gradi di specialismo delle varianti terminologiche che esprimono il concetto di "raccolta differenziata porta a porta"

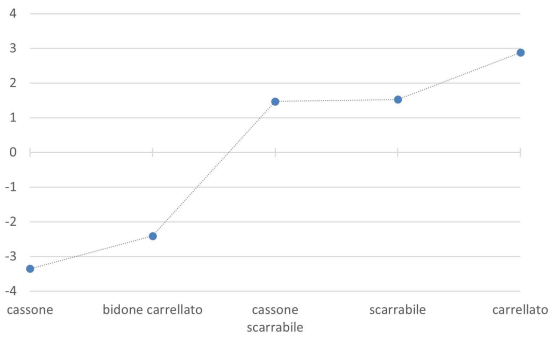


Figura 7: Gradi di specialismo delle varianti terminologiche che esprimono i concetti di “cassone scarrabile” e “bidone carrellato”