NextGenerationEU – **DM 351/2022, CUP I61I22000310007,**
Missione 1 Componente 3 "Turismo e Cultura 4.0"

UNIVERSITÀ DEGLI STUDI DI NAPOLI "PARTHENOPE"

SCUOLA INTERDIPARTIMENTALE DI ECONOMIA E GIURISPRUDENZA

Dipartimento di Studi Economici e Giuridici (DiSEG)

Corso di Dottorato in
Studi Linguistici, Terminologici e Interculturali
XXXVIII Ciclo

Tesi di dottorato in SC:
10/H1-LINGUA, LETTERATURA E CULTURA FRANCESE

CORPUS E STRUMENTI LINGUISTICI
PER LA PROTEZIONE E LA VALORIZZAZIONE TURISTICA
DEL PATRIMONIO NATURALE MARINO

CORPUS ET OUTILS LINGUISTIQUES
POUR LA PROTECTION ET LA VALORISATION TOURISTIQUE
DU PATRIMOINE NATUREL MARIN

TUTOR
Chiar.ma Prof.ssa
Silvia Domenica Zollo

CANDIDATA
Dott.ssa Virginia Carrella
MATR. DR14000004

COORDINATRICE
Chiar.ma Prof.ssa
Maria Giovanna Petrillo

ANNO ACCADEMICO 2024/2025

Alle mie colleghe ed ai miei colleghi,
alle dottorande ed ai dottorandi che ho incrociato nel mio cammino,
ai giovani, luce inquieta e linfa vitale di ogni società,
dedico questo lavoro, maturato nella passione, nella tenacia e nel respiro plurale
che ho colto nei vostri sguardi, ogni giorno.
In ciascuno di voi ho riconosciuto l'eco di un pensiero critico, la forza discreta
della curiosità e la certezza che la conoscenza vive solo dove arde il dialogo.

Venca dunque la perseveranza,
perché, se la fatica è tanta, il premio non sarà mediocre.
Tutte le cose preziose son poste nel difficile.
Giordano Bruno, La cena de le ceneri, 1584

Indice

INTRODUZIONE

Il presente elaborato di tesi nasce nell'ambito della borsa di dottorato del corso in "Studi Linguistici, Terminologici e Interculturali" (XXXVIII ciclo), a valere sui fondi del Piano Nazionale di Ripresa e Resilienza (PNRR)[1], Missione 1, Componente 3 "Turismo e Cultura 4.0"[2], *NextGenerationEU* dell'Unione Europea[3].

Tale finanziamento si inserisce nelle finalità previste dall'Investimento 1.1 "Strategia digitale e piattaforme per il patrimonio culturale", volto alla digitalizzazione, valorizzazione e fruizione accessibile del patrimonio culturale e naturale, finalità che – a loro volta – si stanziano nel più ampio orizzonte degli obiettivi di sviluppo sostenibile promossi dall'Agenda 2030 delle Nazioni Unite[4], in particolare del Goal 14, "Vita sott'acqua", dedicato alla conservazione ed all'uso durevole degli oceani, dei mari e delle risorse marine.

Lo studio, infatti, si colloca nel quadro del progetto di ricerca *Ocean Literacy*[5], attivo presso l'Università degli Studi di Napoli "Parthenope" dal 2022, che nasce con un duplice obiettivo: da un lato, raccogliere, classificare ed analizzare lessici e discorsi tecno-scientifici nel campo della biologia marina, sperimentando approcci teorici e metodologici avanzati nell'ambito della linguistica dei corpora, della terminologia e della lessicografia specialistica; dall'altro di valorizzare il patrimonio naturale marino, specialmente in ottica di promozione turistica del territorio, attraverso la creazione di risorse linguistiche multilingue che intendono rispondere alle esigenze di cittadini, viaggiatori, operatori e guide turistiche ed ambientali, divulgatori scientifici e professionisti del settore, offrendo strumenti affidabili, controllati ed accessibili per la diffusione dei saperi.

---

[1] *Ibidem.*

[2] PNRR, M1C2 "Patrimonio culturale per la prossima generazione*", Missione 1 Componente 3 Turismo e Cultura*, pp. 97-107, 2022.

[3] Ai sensi del DM 351/2022.

[4] Nazioni Unite, Goal 14 "Vita sott'Acqua", *Agenda 2030*, (link: https://unric.org/it/obiettivo-14-conservare-e-utilizzare-in-modo-durevole-gli-oceani-i-mari-e-le-risorse-marine-per-uno-sviluppo-sostenibile/).

[5] Zollo S. D., «Lexiques et corpus au service de la littératie océanique : propriétés et relations lexicales dans le domaine de la faune marine», *Studia Universitatis Babes-Bolyai - Philologia*, 2024a.

In particolare, nel presente elaborato ci proponiamo di illustrare le scelte ed i processi che hanno condotto all'elaborazione delle risorse lessicografiche: nel primo capitolo ci siamo occupati di definire e delimitare il dominio di ricerca in ottica terminologica ed in rapporto con il concetto di Dominio in Biologia marina; nei capitoli 2 e 3 abbiamo poi stabilito i fondamenti teorici e metodologici del lavoro, descrivendo, in primo luogo, le fasi di costituzione del corpus bilingue *ZooCor* (fr-it)[6] e di sviluppo del software *CorpusBuilder* (sviluppato durante il soggiorno di ricerca presso il laboratorio di informatica dell'Université de Pau et des Pays de l'Adour). In secondo luogo, abbiamo delineato il quadro teorico di riferimento che si propone di sperimentare l'uso della Lessicologia Esplicativa e Combinatoria (LEC)[7] e delle Funzioni Lessicali (FL), al fine di rendere il lessico specialistico accessibile anche per categorie di utenti non esperti, come nel caso dei fruitori di una strategia di promozione turistica territoriale.

Infine, nei capitoli 4 e 5 abbiamo illustrato l'applicazione dell'impianto teorico-metodologico, partendo dalla descrizione delle fasi di estrazione terminologica automatica dal corpus, per concludere con l'esposizione dei risultati di una modellazione delle relazioni semantico-lessicali dei termini estratti condotta secondo i principi della LEC.

In tal modo, il presente elaborato si propone di contribuire all'implementazione di un modello per la costruzione di una terminologia condivisa e multilingue nel campo del patrimonio naturalistico marino, fruibile dalla comunità scientifica, dai professionisti della comunicazione ambientale e dagli operatori turistici, attraverso nuove forme di mediazione linguistica coerenti con i principi di *Open Science*, oltre che con gli obiettivi di transizione digitale e verde che guidano il PNRR e l'Agenda 2030. La terminologia diventa così un veicolo privilegiato per promuovere un dialogo interdisciplinare tra lingua, scienze naturali e comunicazione specialistica, in ottica di valorizzazione del territorio.

[6] Zollo S.-D., *ZooCor Corpus (fr-it) (First version)*, [Data set], Zenodo, (disponibile online *https://doi.org/10.5281/zenodo.11210074*), 2024b.

[7] Mel'čuk I., *et al*., *Introduction à la lexicologie explicative et combinatoire*, Duculot, Louvain-la-Neuve 1995.

CAPITOLO 1. DEFINIZIONE DEL DOMINIO IN BIOLOGIA MARINA

Abstract

Il primo capitolo definisce e delimita il dominio della tesi, un passaggio essenziale per l'analisi di un linguaggio specialistico. Partendo da una rassegna della letteratura, viene esaminato il concetto di dominio in terminologia, le sue diverse accezioni e le sfide legate alla sua delimitazione, oltre che la rispettiva interdipendenza con il termine per la contestualizzazione dei significati specialistici. Il capitolo identifica il dominio della biologia marina attraverso un approccio interdisciplinare che include e privilegia la consultazione degli esperti, al fine anche di integrare la prospettiva terminologica con la classificazione tassonomica. Di conseguenza, viene delineata la struttura gerarchica del dominio relativo alla fauna marina in biologia marina, ovvero il "Regno Animale", che parte dal Dominio Eucarioti fino a giungere al livello tassonomico delle varie "Famiglie" che ne fanno parte: queste ultime costituiranno il punto di partenza per una sotto-categorizzazione in sottodomini di riferimento per le conseguenti fasi di raccolta dei testi e di analisi terminologica.

1.1. Definizioni ed approcci all'analisi del dominio

Il dominio, secondo quanto suggerito da de Bessé, può essere definito, come una « structuration des connaissances »[8], ovvero un'organizzazione coerente delle conoscenze, che consente di identificare, delimitare e denominare una specifica struttura cognitiva, formata da un sistema di concetti interrelati, strettamente connessi.

In particolare, nonostante l'ISO, l'Organizzazione Internazionale di Normalizzazione, definisca il dominio come « branche spécialisée de la connaissance » in cui « les limites […] sont définies selon un point de vue particulier lié à l'objectif visé »[9], al giorno d'oggi le differenti denominazioni dei domini testimoniano una evoluzione, ma soprattutto una difficoltà, nella definizione stessa di dominio. Quali sono le cause di tali problematiche? I domini sono considerati come altamente specializzati e di conseguenza vengono separati da confini che risultano essere molto sottili, in quanto queste delimitazioni tra domini sono spesso soggette anche alle lingue, alle tradizioni ed alle culture di riferimento, a punti di vista specifici, a pratiche sociali ed a bisogni degli utenti.

Ne consegue che solo attraverso la definizione della nozione di dominio è possibile delineare e comprendere la complessità di un sistema concettuale, fornendo così un quadro strutturato e definito delle conoscenze in esso contenute.

In realtà, l'analisi del dominio può essere condotta attraverso diversi approcci, ad esempio attraverso una prospettiva puramente cognitiva, indagando le relazioni tra i concetti all'interno di un dominio, e quindi la loro interdipendenza che struttura la conoscenza[10], ma anche in un'ottica comunicativa e terminologica.

In primo luogo, da un punto di vista comunicativo, il linguista – attraverso uno studio che analizzi le pratiche discorsive e le dinamiche comunicative all'interno di

[8] De Bessé B., « Le domaine », in Béjoint H. et al., *Le sens en terminologie*, Presses Universitaires de Lyon, Lyon 2000, p. 183.

[9] ISO 10241-1:2011 https://www.iso.org/obp/ui/en/#iso:std:iso:10241:-1:ed-1:v1:fr.

[10] In effetti, in molti studi affrontati (incentrati sia propriamente sulla definizione di dominio che sull'analisi delle terminologie specialistiche) la rilevanza della nozione di dominio risiede nella sua capacità di comprendere la complessità dei sistemi concettuali, in quanto attraverso la definizione di un dominio è possibile delineare ed indagare le relazioni tra concetti appartenenti ad uno stesso dominio, essenziali per la comunicazione specialistica.

un dominio – fornisce etichette ai concetti di un sistema nozionale per poter scambiare informazioni, trasmettere nuove conoscenze ed altro, attraverso dati linguistici.

In secondo luogo, nel contesto della prospettiva lessicalista[11] – come riportato da Polguère – appare quanto più appropriato effettuare uno studio del lessico per « champs lexicaux »[12], in quanto la loro identificazione permette di organizzare il lessico specialistico, quindi di raggruppare *lexies* e facilitare così l'analisi lessicografica e la creazione di risorse terminografiche efficienti: in tal senso, il dominio funge da criterio per classificare il lessico specialistico in campi concettuali ben definiti[13].

Infine, da un punto di vista puramente terminologico, gli studi di riferimento offrono diverse prospettive sulla nozione di dominio, che affronteremo nei paragrafi successivi.

1.1.1. Dominio in terminologia: rassegna della letteratura scientifica di riferimento

In terminologia, il concetto di dominio è molto ampio: secondo de Bessé, il dominio è « un ensemble organisé de concepts […] interdépendants, liés entre eux »[14], che – con un ruolo « super-générique »[15] – costituisce una *classe*, ovvero un insieme di oggetti di conoscenza che hanno caratteristiche comuni tra loro.

In quest'ottica, il dominio permette di raggruppare e di ordinare le nozioni e di costruire i sistemi cognitivi, indicando le prospettive da adottare per delimitare e descrivere un concetto. Nello specifico, de Bessé distingue diverse categorie di

[11] Nello specifico, tale attenzione nei confronti della definizione del dominio di riferimento rientra in uno studio – il nostro – che si propone di indagare in maniera lessicografica un lessico specialistico (§ capitolo 5).

[12] Polguère A., *Lexicologie et sémantique lexicale : Notions fondamentales*, Presses de l'Université de Montréal, Montréal 2016; p. 230.

[13] Spesso il campo lessicale è definito in maniera molto generica in terminologia: negli studi è frequente, infatti, la corrispondenza tra questo ed alcune discipline scientifiche o tecniche. (Cfr. L'Homme M.-C., *Lexical semantics for terminology*, John Benjamins Publishing Company, Philadelphia 2020).

[14] De Bessé B., *op. cit*., p. 183.

[15] *Ibidem.*

domini, che si strutturano in base alla loro stessa natura: nell'ambito di un insieme di conoscenze strutturate sulla base di una tematica, egli introduce la nozione di *domaine de connaissances*, ovvero di un « savoir constitué […] selon une thématique »[16], come ad esempio la matematica, la zoologia, l'economia ecc. Viceversa, per distinguere un particolare tipo di discorso, di enunciato, di comunicazione, in un'ottica sociolinguistica de Bessé parla di *domaine de discours*, che, a sua volta, si distingue dalla terza macrocategoria, le *domaine d'activité*, che invece permette di identificare un campo d'azione: « […] il est constitué d'un ensemble de procédés bien définis destinés à produire certains résultats […] », come, ad esempio, i trattati europei « […] conçus en fonction d'un besoin particulier et correspondant à une activité précise »[17]. Sulla base di queste tre macrocategorie, de Bessé distingue diverse tipologie di domini, come ad esempio *les domaines terminologiques*, *les domaines documentaires*, *les domaines sémantiques* e *les domaines terminographiques.* Questa differenza si stanzia sulla concezione dei domini come dei punti di vista « […] délimités en fonction des visions des connaissances, des pratiques sociales et des besoins des utilisateurs »[18]. Tuttavia, tale soluzione non basta a chi rivendica e sottolinea l'interdipendenza dei termini dai domini e la loro specificità all'interno di questi ultimi.

Nel nostro studio ci siamo basati sulla definizione di dominio terminologico, che de Bessé adotta per descrivere i domini costruiti « *a priori*, notamment par les scientifiques et les juristes »[19] e rappresentati da tutti quei sistemi concettuali risultanti « […] de l'organisation des connaissances, de l'élaboration de systèmes cognitifs, de la construction de théories, de la constitution et du réglage des savoirs […] »[20], avvicinandosi alla classificazione che viene effettuata nelle materie scientifiche (come nel nostro caso in biologia), in cui la distinzione tra un dominio e l'altro è più netta.

In parallelo, Delavigne sottolinea la necessità di una chiara struttura nozionale per definire un dominio; tuttavia, nonostante il concetto di dominio fornisca un quadro

[16] *Ivi*; p. 184.
[17] *Ivi;* pp. 184-185.
[18] *Ivi;* p. 187.
[19] *Ivi;* p. 188.
[20] *Ibidem.*

utile per comprendere le relazioni concettuali e la specializzazione all'interno di un campo specifico e venga considerato come « Principe de base de la reconnaissance des termes scientifiques et techniques, […] un des piliers de la théorie terminologique et, notamment, des études de néologie »[21], l'autrice ne evidenzia la fragilità epistemologica, legata proprio alla difficoltà di circoscriverne i confini e di adattarla a diversi contesti:

> Étant donné sa centralité, il est donc nécessaire de s'assurer de sa valeur théorique. Car, si cette notion de sens commun est assurément utile, elle se révèle, malgré sa commodité, inadaptée pour penser les pratiques discursives dans leur complexité. [...] Le flou de la notion et la difficulté de la circonscrire la rattachent en effet à une axiomatique et diminuent fortement sa valeur épistémologique.[22]

Infatti, Delavigne sostiene che al giorno d'oggi « […] ce qui fécondent les sciences et les techniques […] c'est l'échange, l'hybridation, le métissage »[23], di conseguenza vi è una « […] interpénétration constante […] »[24] dei saperi, giustificata dall'accelerazione dello sviluppo della conoscenza e dell'interdisciplinarietà che hanno contribuito ad « […] éclater les domaines […] »[25], « […] au point qu'il semble nécessaire de penser la notion de domaine différemment »[26].

Dinanzi ad una vera e propria crisi del dominio, Delavigne suggerisce una visione di esso come uno spazio aperto, come un « […] lieu de convergence de pratiques […] »[27]; l'autrice parla di « réseaux de nœuds »[28] di saperi, di pratiche, sancendo l'interdipendenza di tutte le scienze e tutte le tecniche tra loro: un'intima tessitura di discipline, oggetti, dove le parole circolano in schemi, abitudini, modi

[21] Delavigne V., « La notion de domaine en question – À propos de l'environnement », in Gérard C., Balnat V., *Neologica – Néologie et environnement*, n. 16, Éditions Didascaliques, 2022, p. 1.
[22] *Ibidem.*
[23] Delavigne V., « Le domaine aujourd'hui. Une notion à repenser », in Candel D., *Le traitement des marques de domaine en terminologie*, Cahiers du LCPE, Paris 2002; p. 9.
[24] *Ibidem.*
[25] *Ibidem.*
[26] *Ibidem.*
[27] *Ivi;* p. 10.
[28] *Ibidem.*

di pensare, ragionamenti, che a loro volta fecondano le scienze. Tuttavia, Delavigne specifica che non si tratta di una progressiva complicazione della nozione di dominio, bensì di una sua integrazione nella sua stessa costruzione: « […] c'est ce que la notion de domaine, ainsi complexifiée, se doit d'intégrer […] »[29].

Questa moltitudine di saperi interdipendenti tra loro introduce un'ulteriore visione, che garantisce una trasversalità dei domini: l'*arbre de domaine* che Tremblay e Rondeau definiscono come una rappresentazione sincronica di una risorsa nozionale. Tale figura rappresenta « […] un état de langue à un moment de son évolution […] », che tuttavia non deve essere considerato come un « système figé »[30], bensì è l'immagine di una fonte nozionale che si trasforma al bisogno, secondo i cambiamenti del dominio. Spesso l'*arbre de domaine* viene messo in relazione con la nozione di tesauro, che ha come obiettivo la catalogazione ed il reperimento di informazioni contenute nei documenti, al contrario del primo che è utilizzato per indagare la terminologia stessa di un dominio. Inoltre, un tesauro può molte volte rispondere ai bisogni documentari di più domini e differisce dall'*arbre de domaine* soprattutto nella sua rappresentazione: secondo il sito ufficiale dell'Unione Europea, il tesauro rappresenta "un vocabolario controllato e strutturato i cui concetti sono rappresentati da etichette (label)"[31], è dunque uno strumento di controllo del vocabolario che mostra le relazioni tra i termini, a differenza dell'albero di dominio che invece rappresenta una struttura gerarchica che organizza le informazioni all'interno di un dominio specifico. Al contempo, quindi, un dominio viene rappresentato da una rete di concetti interconnessi, visualizzati ad esempio tramite mappe concettuali.

Da un punto di vista lessicografico, Candel nel 1979 propone una disamina dei tre principali dizionari della lingua francese al fine di rivelare una notevole eterogeneità sia nella presentazione che nella denominazione dei domini. In primis,

---

[29] *Ibidem.*

[30] Tremblay D., Rondeau G., *La notion d'arbre de domaine appliquée à la terminologie comme discipline*, extrait de la thèse de maîtrise présentée en 1982 et préparée sous la direction de G. Rondeau, professeur titulaire de terminologie au Département de langue et linguistique de l'Université Laval, texte abrégé des pages 1 et 8 à 23, 1982; p. 28.

[31] https://op. europa.eu/it/web/eu-vocabularies/thesauri#:~:text=Un%20thesaurus%20%C3%A8%20un%20vocabolario,rappresentati%20da%20etichette%20(label). Consultato nel 2025.

Candel analizza il dizionario *Le Grand Robert (GR)*[32] che su circa 609 voci principali, individua un insieme di 70 domini, presentati con caratteri tipografici diversi da ciò che precede o segue, di cui circa 8 vengono introdotti dalla formula « en termes de… », che indica un utilizzo della parola in situazioni più particolari:

> […] elle marque une utilisation du mot évoqué dans une situation ou un ensemble de situations particulières, par les spécialistes d'une profession ou les membres d'un groupe particuliers. Cette formule implique une précision au niveau du discours, de la langue ; il s'agit explicitement du signe.[33]

In secundis, Candel riporta che nel *Le Grand Larousse De La Langue Française (GLLF)*[34] invece, su circa 657 voci principali, si rilevano 26 domini, perfettamente integrati nel corpo della definizione, senza particolari artifici tipografici, bensì introdotti dalla preposizione « en » ed esplicitati adottando numerose localizzazioni di attività di spazio o di tempo:

> Certaines formules du GLLF sont, d'autre part, tout à fait différentes. Ce sont les localisations qui introduisent les définitions, localisations (a) dans une activité ou un ensemble d'activités, (b) dans l'espace, ou (c) dans le temps.[35]

Inoltre, spesso, il GLLF inserisce il dominio nel contenuto stesso della definizione. Infine, l'autore analizza un terzo dizionario che, rispetto al GLLF ed al GR, semplifica la nomenclatura dei domini: il *Trésor de la Langue Française (TLF)*[36] presenta la maggior parte dei domini in forma semplice, senza preposizioni « en » o « dans », ma messi in evidenza dalla loro rappresentazione in corsivo, in maiuscolo, con la specifica dei sottodomini tra parentesi o, addirittura, preceduti dall'indicazione « domaine de… ».

---

[32] Le Grand Robert, tome I, 1953.
[33] Candel D., « La présentation par domaines des emplois scientifiques et techniques dans quelques dictionnaires de langue », in Delesalle S., Rey A., *Langue française*, vol. 43, Larousse, Paris 1979, pp. 100-115; p. 102.
[34] Grand Larousse de la langue française, tome I, 1961.
[35] Candel, *op.cit.*, p. 103.
[36] Trésor de la Langue Française, tome IV, 1975.

Lo stesso avviene nei dizionari elettronici, dove vengono adoperati delle « marques de niveau de langue », delle etichette simboliche per indicare l'impiego di una parola in un dominio:

> *techn.* : mot du langage technique ; *admin.* : mot employé dans la langue écrite de l'administration ; *didact*. : terme de la langue savante, que l'on peut trouver dans un traité, dans un cours, mais que l'on n'utilise pas dans le langage courant.[37]

Tuttavia, secondo l'indagine condotta da Buvet e Colas, l'applicazione di tali *marques* non è sempre coerente, in quanto diversi dizionari elettronici possono trattare lo stesso termine in modo diverso; ciò evidenzia, ancora una volta, la difficoltà di definire e di classificare i domini in modo univoco, anche in lessicografia, nei dizionari elettronici. In particolare, però, nella strutturazione stessa di un dizionario elettronico, tra le varie proposte risolutive che gli autori riportano nell'identificazione e nella classificazione dei domini, spicca la definizione data da Quemada, il quale si dirotta maggiormente sull'utilitarismo intrinseco del termine: « […] les indications de domaine n'indiquent pas seulement le champ d'expérience dont relève le mot (cf. agentif, ling.) mais aussi à propos duquel on l'utilise »[38].

Ne consegue che sia la lessicografia che la terminologia si occupano di definire e classificare i domini, ma siccome in terminologia vi è una forte propensione verso la necessità di specificare i domini, Buvet e Colas persistono con la centralità di tale concetto come « constitutive de la terminologie »[39].

In realtà, tale importanza viene ulteriormente rimarcata anche da altri studiosi, come osserveremo nella definizione del rapporto tra il dominio ed i termini, nel paragrafo

[37] Rey-Debove et Bellefonds 1989, in Buvet P.-A., Mathieu-Colas M., « Les champs 'domaine' et 'sous-domaine' dans les dictionnaires électroniques », in *Cahiers de Lexicologie*, vol. 75, Classiques Garnier, Paris 1999, p. 173.

[38] Quemada *et al*. 1984, in Buvet P.-A., Mathieu-Colas M., « Les champs 'domaine' et 'sous-domaine' dans les dictionnaires électroniques », in *Cahiers de Lexicologie*, vol. 75, Classiques Garnier, Paris 1999, p. 175.

[39] Buvet P.-A., Mathieu-Colas M., « Les champs 'domaine' et 'sous-domaine' dans les dictionnaires électroniques », in *Cahiers de Lexicologie*, vol. 75, Classiques Garnier, Paris 1999, p. 174.

successivo; al contempo Estopà[40] mette in evidenza il ruolo del dominio nella specializzazione del significato terminologico, dominio che, a sua volta, viene descritto da Taifi[41] come uno spazio chiuso con centro, frontiera e periferia; del resto, come denota L'Homme, la nozione di "special subject field (or domain)"[42] prevede anche la delimitazione e la classificazione del dominio da parte di chi ne effettua un'analisi, stabilendone i confini entro i quali i terminologi considerano i termini in funzione della loro rilevanza, o meno, al dominio.

1.1.2. Dominio in terminologia: rapporto tra dominio e termine

Secondo de Bessé, il dominio, insieme con il concetto e la definizione, costituisce uno dei tre pilastri su cui si basa il termine, e che quindi « Le terme se caractérise par son appartenance à un domaine. Le domaine fait donc partie des informations qui doivent obligatoirement accompagner le terme »[43].
Di conseguenza, il dominio fornisce il contesto necessario alla determinazione del significato di un termine, sancendo un'interdipendenza essenziale tra i due, che egli valuta come uno dei principali tratti distintivi con la "parola":

> Pour exister en tant que terme, une forme linguistique doit désigner un concept appartenant à un domaine et déterminé par une définition. Le domaine représente l'un des trois éléments du trépied sur lequel repose le terme (les deux autres étant le concept et la définition) […]. Le concept, sa définition (et son terme) appartiennent obligatoirement a un domaine. […] C'est notamment l'appartenance à un domaine qui permet de distinguer le terme et le mot. C'est également le plus souvent grâce au domaine que l'on peut distinguer les termes homonymes.[44]

[40] Estopà Bagot R., « Les unités de signification spécialisées élargissant l'objet du travail en terminologie », in Kageura K., Temmerman R., *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, vol. 7, n. 2, John Benjamins Publishing Company, 2001, pp. 217-237.
[41] Taifi M., *Sémantique linguistique. Référence, prédication et modalité*, Fès, UFR : Sciences du Langage, 2000.
[42] L'Homme 2020, *op. cit*., p. 37.
[43] De Bessé B. 2000, *op. cit*., p. 190.
[44] *Ivi;* p. 182.

Nei linguaggi specialistici il dominio permette di contestualizzare il significato dei termini, e, a sua volta, risulta strettamente legato al loro campo di utilizzo: infatti come afferma Slodzian « le domaine est au terme ce que le contexte est au mot »[45] e che – quindi – come specifica Delavigne « le terme à l'intérieur d'un domaine » assume « un certain nombre de caractéristiques » tali da distinguerlo dal suo utilizzo nella « langue courante »[46].

Ne consegue la necessità di definire tale contesto, che fa capo alla questione riguardo la complessità intrinseca nella definizione stessa del termine, sorta nel momento in cui i linguisti, piuttosto che gli specialisti del settore, si sono occupati della produzione di dizionari specialistici. Tra le divergenti teorie esposte fino ad oggi, emerge quella in cui Estopà parla di una « unité de signification spécialisée »[47], introducendo ed anticipando la denominazione di "unità terminologica"[48], che sottende il suo stesso ruolo di contenitore di un significato la cui specializzazione fa sì che, come enunciato da Frassi, « un spécialiste du domaine, un traducteur, un rédacteur et un journaliste peuvent avoir des vues divergentes sur ce qui constitue un terme, surtout multilexémique »[49].

L'unità terminologica – per assumere il significato più opportuno – subisce una contestualizzazione in micro e/o macro *strutture*, che possiamo identificare nei domini, i quali, secondo Delavigne, sono costruiti « à partir d'un système de concepts […] ou de notions » [50]; questi ultimi nel definire e nel caratterizzare i domini « font partie d'une structure notionnelle, et c'est cette structure notionnelle qui est censée représenter le domaine »[51].

Dal nostro punto di vista, è essenziale considerare come le unità terminologiche si evolvano e si adattino alle nuove esigenze ed ai cambiamenti di tali strutture

[45] Slodzian M., « Comment revisiter la doctrine terminologique aujourd'hui ? », *La Banque des Mots*, n. spécial, 1995, pp. 11-18; p. 14.
[46] Delavigne V. 2002, *op. cit*., p. 2.
[47] Estopà Bagot R. 2001, *op. cit*., p.1.
[48] Da ora in poi, faremo riferimento al termine come "unità terminologica" (L'Homme 2004, Frassi 2019, *et al*.)
[49] Frassi P., Humbley H., Calvi S., « Fouille de textes et repérage d'unités phraséologiques », in Roche C., *Terminologie & ontologie : Théories et applications*, Actes de la conférence TOTh, Presses universitaires Savoie Mont Blanc, Chambéry 2019, p. 324.
[50] In questo senso, la terminologia ha a lungo utilizzato l'uno in modo interscambiabile con l'altro, anche se il concetto tende attualmente a prendere piede.
[51] Delavigne V. 2002, *op. cit*., pp. 6-7.

nozionali nei contesti di utilizzo. L'evoluzione della terminologia, infatti, è strettamente legata all'evoluzione del dominio stesso, il quale può essere, ancora di più, esposto all'influenza di fattori socioculturali, tecnologici e scientifici.

Inoltre, la riflessione sulla definizione e sull'identificazione delle unità terminologiche – ed in particolare il processo di determinare ciò che costituisce un'unità terminologica rilevante – richiede un approccio interdisciplinare ed olistico che coinvolga linguisti, specialisti del settore e altri attori coinvolti nella produzione e nell'uso dei termini specializzati[52]. Un'analisi approfondita delle strutture concettuali e linguistiche – dall'interno ed all'interno di un dominio specifico – può contribuire ad una migliore comprensione delle unità terminologiche e dei loro rapporti nel contesto di utilizzo, essenziale per lo sviluppo di risorse terminografiche efficaci e per facilitare la comunicazione specialistica tra gli esperti del settore ed un pubblico anche di non esperti.

In tale contesto specialistico, si sviluppa l'analisi di una particolare tipologia di dominio, il dominio tecnico-scientifico.

1.2. Definizione ed approcci all'analisi del dominio tecnico-scientifico

Nel campo della terminologia, il concetto di dominio tecnico-scientifico gioca un ruolo fondamentale nel processo di analisi – e quindi di definizione e comprensione – dei termini. Come riportano Bouveret e Gaudin, il lavoro del terminologo implica una riflessione sui termini a partire dai discorsi « qui sont liés à des pratiques socialisées, pour aller vers la langue et non l'inverse »[53].

Il significato stesso di dominio, come suggerisce Candel « sert à marquer la répartition de l'expérience humaine en secteurs »[54] che si riferisca al referente o al concetto, e conduce con sé anche informazioni sul livello del linguaggio e sul discorso utilizzato. La suddivisione in domini può essere basata su criteri semantici, legati al concetto e alla classe di oggetti cui il termine si riferisce, oppure – come

[52] Questa considerazione è alla base del nostro processo di analisi linguistica di specialità, che verrà affrontata nei paragrafi successivi (§ capitolo 3).

[53] Bouveret M. & Gaudin F., « Du flou dans les catégorisations : le cas de la bioinformatique », in De Schaetzen C., *Terminologie et interdisciplinarité*, Peeters Publishers, Louvain 1997, p. 65.

[54] Candel D.1979, *op. cit*., p. 100.

riporta Depecker – su criteri pragmatici, indicando un legame tra l'uso del termine ed il contesto d'uso:

> Ces domaines sont la projection terminologique de secteurs d'activité, de champs de connaissance, d'objets du monde, etc. dans cette reconstruction terminologique, la notion de *spécialisation* est déterminante, car elle contribue à caractériser l'unité terminologique.[55]

In particolare, Candel – teorizzando la classificazione dei domini tematici, che corrispondono a loro volta a settori organizzati di conoscenza – distingue i domini teorici ed i domini tecnici:

— i domini teorici, come la filosofia e le scienze, si riferiscono a « des structurations notionnelles, à des théories différentes engendrant des termes qui véhiculent, quant au même objet, des valeurs différentes »[56];
— i domini tecnici, invece, riguardano situazioni concrete e contingenti pratiche, come progetti, strumenti e attività volte a modificare il contesto circostante. La definizione chiara dei domini non è sempre facile, soprattutto per alcuni domini complessi, come la medicina, la terapia, la chirurgia e le tecnologie specializzate.

A tal proposito l'autrice suggerisce come gli autori di dizionari linguistici dovrebbero analizzare attentamente le distinzioni dei domini specialistici e le loro implicazioni « au niveau du signe, du mot, du discours »[57], al fine di stabilire le relazioni tra la struttura tematica, i livelli di linguaggio ed i tipi di utilizzo.

Tuttavia, la definizione dei domini rimane un'operazione complessa, evidenziando la necessità di ulteriori studi e collaborazioni tra esperti per evitare incoerenze e ambiguità. Quemada propone come soluzione che le indicazioni sul dominio non indichino « seulement le champ d'expérience dont relève le mot […] mais aussi à

[55] Depecker L., « Contribution de la terminologie à la linguistique », in *Langages*, n. 157, Armand Colin, Paris 2005, pp. 6-13; p. 6.
[56] Candel D. 1979, *op. cit*., p. 100.
[57] *Ivi*; p. 101.

propos duquel on l'utilise »[58]: è fondamentale, infatti, avvicinarsi il più possibile all'ambiente ed al dominio in questione, al fine di aumentare la conoscenza degli usi e le principali nozioni della terminologia di riferimento.

1.3. Identificazione del dominio di ricerca

Una volta riportate le definizioni di dominio, di dominio tecnico-scientifico e la loro proiezioni nei nostri studi linguistici, ci siamo interrogati sull'identificazione e sulla delimitazione del dominio all'interno del nostro progetto[59].

1.3.1. La consultazione degli esperti del settore

Come suggerito da Auger *et al*.,

> Il serait impensable de se lancer dans un travail terminologique spécialisé sans avoir acquis au préalable des connaissances générales sur le sujet.[60]

Infatti, prima di procedere alla definizione del dominio di ricerca, ci è sembrato necessario interagire con gli esperti del settore. In particolar modo, ci siamo ispirati anche ai modelli di chi[61] ha ritenuto l'interazione con gli esperti del settore, indispensabile per tutte le applicazioni e/o gli studi necessari per la consultazione delle banche dati dedicate alla descrizione degli elementi del campo di studio in oggetto.

---

[58] Quemada *et al.* in Buvet P.-A., Mathieu-Colas M., « Les champs 'domaine' et 'sous-domaine' dans les dictionnaires électroniques », in *Cahiers de Lexicologie*, vol. 75, Classiques Garnier, Paris 1999, pp. 173-191.

[59] Per fare ciò abbiamo previsto tre passaggi fondamentali che andremo a sviluppare nei successivi paragrafi, ovvero una prima consultazione con gli esperti del nostro settore di riferimento, la definizione del dominio nel nostro settore di riferimento e la sua strutturazione in ulteriori sottodomini.

[60] Auger P. & Rousseau L.-J., *Méthodologie de la recherche terminologique*, Office de la langue française, Service des travaux terminologiques, L'Éditeur officiel du Québec, 1978, pp. 15-16.

[61] Ad esempio, d'ispirazione per il nostro studio è stata l'analisi nel campo della bioinformatica da parte degli autori Bouveret & Gaudin (1997).

Ma chi sono gli esperti? Come suggerisce Delavigne, un esperto, un professionista del settore, « est toujours un spécialiste de quelque chose »[62] in altre parole, non è mai uno specialista in senso assoluto, ma sempre rispetto a una precisa area del sapere o campo di ricerca.

Se, quindi, l'obiettivo consiste nel comprendere e divulgare le conoscenze di un dominio scientifico, perché coinvolgere i linguisti in un tale contesto? La ragione è semplice e circostanziale. Partendo dallo studio delle irregolarità nelle pratiche linguistiche, delle difficoltà nell'utilizzo di lingue differenti e dei diversi registri nei quali si esprimono i parlanti del contesto scientifico (*i.e.* dizionario delle bio-industrie[63]; vocabolario dell'ingegneria genetica[64]) ci è sembrato pertinente gestire la ricerca anche come fonte di nuove sinergie.

Il nostro intervento, infatti, nasce dalla quasi-assenza di banche dati che sintetizzino informazioni sul mondo marino, con conseguenti effetti anche socioeconomici[65]. Tuttavia, nel caso particolare di un "non-esperto" ad esempio, sarà necessario familiarizzare sia con il linguaggio usato per descrivere un determinato dominio, sia con il dominio in sé: è evidente quindi quanto sia necessario acquisire sia la conoscenza linguistica sia una conoscenza concettuale. Se da un lato la prima mira ad una assimilazione di tutte le strutture linguistiche, termini specialistici, collocazioni *etc.*, la seconda si concentra sulla profonda consapevolezza dei termini utilizzati, per l'edificazione della piena, o quasi, padronanza dei concetti celati oltre i termini. Tale conoscenza di pratiche, usi ed avvenimenti concreti che creano l'esperienza vera di un dominio è ritracciabile anche attraverso l'interfacciarsi con le realtà che si desidera descrivere, non soltanto tramite lo studio di esse.

Questa situazione ci ha motivate ad interagire con esperti (come biologi, volontari, *etc.*), per tentare di ottimizzare gli scambi e la collaborazione tra gli studiosi, i divulgatori e gli utenti di future banche dati digitali (dizionari, tesauri, ontologie ecc.), come riportato da Zollo:

---

[62] Delavigne V. 2002, *op. cit*., p. 8.

[63] Bouveret M. & Gaudin F., 1997, *op. cit*., pp. 63-72.

[64] Guespin L., « La circulation terminologique et les rapports science-technique-production », in *Cahiers de linguistique sociale*, n. 18, 1991, pp. 59-80.

[65] L'obiettivo è studiare e proporre soluzioni per mitigare, se non risolvere, le difficoltà di comunicazione emerse dalla collaborazione tra biologi, turisti o semplici divulgatori della biologia marina (Zollo 2024a).

être capable de divulguer ces savoirs, dans le contexte scientifique comme vis-à-vis du grand public, demande la connaissance d'un langage technique qui évolue constamment et qui s'adapte aux découvertes et aux changements de notre planète.[66]

Fig. 1.1. Incontri con gli esperti: visite all'acquario della Stazione Zoologica di Napoli.

Pertanto – in tale scenario – il ruolo del terminologo è fondamentale: come suggeriscono Auger e Rosseau, egli deve osservare e mettersi costantemente in contatto con il mondo reale, per captarne le strutture su cui si riflettono le attività concrete (Fig. 1.2.):

> […] l'attention du terminologue doit porter sur la structuration des notions relatives à des objets du monde réel (*champ notionnel*). Cette structuration est souvent le reflet des activités concrètes propres à un domaine donné.[67]

dunque, come riporta Zanola, il terminologo « se présente comme l'intermédiaire qui porte, à la connaissance du grand public, les nouvelles dénominations »[68].

[66] Zollo S. D. 2024a, *op. cit*., p. 231.
[67] Auger P. & Rousseau L.-J., 1978, *op. cit*., p. 17.
[68] Grimaldi C., Marzi E., Puccini P., Zanola M., Zollo S. D. (a cura di), *Terminologia e interculturalità. Problematiche e prospettive*, I libri di Emil, Reggio Emilia, 2022; p. 59.

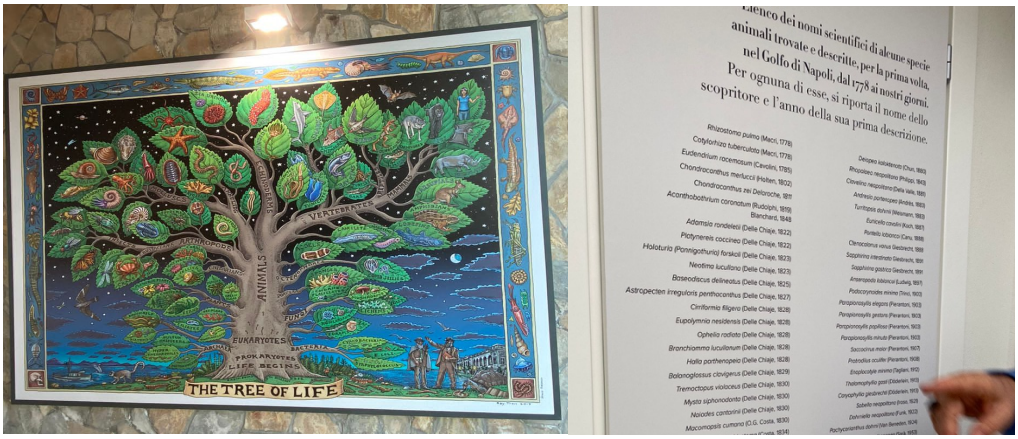


Fig. 1.2. Incontri con gli esperti: visite al museo Darwin Dohrn di Napoli, albero tassonomico delle specie, lista dei nomi scientifici di specie animali trovate e descritte per la prima volta del Golfo di Napoli.

1.3.2. Definizione di dominio in biologia marina

D'altro canto, in biologia marina, il Dominio può essere definito come un *taxon*[69], cioè un gruppo gerarchicamente superiore agli altri[70], in cui gli organismi "hanno certe caratteristiche in comune e si ritiene abbiano un antenato comune"[71] ed il cui livello tassonomico si sviluppa come segue:

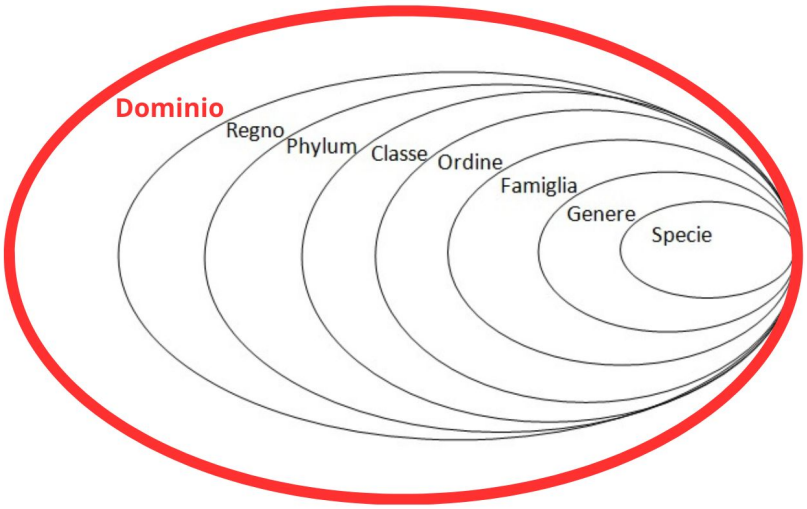


Fig. 1.3. Categorie tassonomiche in biologia (ispirato a Sandulli 2011, p. 97).

[69] "Altri sistemi riconoscono un nuovo *taxon*, il dominio, che è gerarchicamente più alto del regno". Sandulli R., *Biologia marina*, McGraw-Hill, Milano 2011, p. 96.
[70] In tal senso si motiva l'uso del termine come nome proprio, con la prima lettera maiuscola.
[71] Sandulli R. 2011, *op. cit.*, p. 95.

Infatti, nelle scienze biologiche i Domini sono stabiliti *a priori*, attraverso una classificazione oggettiva e non modificabile:

> La démarche scientifique est par essence classificatoire. Ainsi, la zoologie classe les animaux en familles, en genres et en espèces selon les principes suivants : chaque individu ne peut appartenir qu'à une seule classe ; aucune classe ne doit rester vide ; tous les individus doivent trouver place dans une classe. Les taxinomies proposent un ordre à l'intérieur duquel sont rangés les individus ou les objets.[72]

Nel contesto della biologia, il concetto di Dominio deriva dall'approccio tassonomico: fa parte di una gerarchia che va dal livello più generale a quello più specifico, consentendo ai biologi marini di comprendere e categorizzare la vasta diversità della vita marina.

Per comprendere appieno questa dinamica, possiamo far riferimento alla tassonomia standard in biologia, che si articola attraverso una serie di livelli gerarchici, comunemente noti come Domini, Regni, Phyla, Classi, Ordini, Famiglie, Generi, Specie e Sottospecie (Fig.1.4.): questi livelli consentono agli esperti di organizzare e classificare gli organismi in base alle loro caratteristiche morfologiche, genetiche ed ecologiche[73].

Livello	**Esempio in italiano**
Dominio	Eucarioti
Regno	Animale
Phylum	Cordati
Classe	Mammiferi
Ordine	Primati
Famiglia	Umanoidi
Genere	Homo
Specie	Homo sapiens

[72] De Bessé B. 2000, *op. cit*., p. 188.
[73] Sandulli R. 2011, *op. cit*.

Livello	Esempio in italiano
Sottospecie	sapiens

Fig. 1.4. Categorie tassonomiche ed esempi (ispirato a Sandulli 2011, p. 97).

Nella biologia marina, il Dominio potrebbe includere organismi come le piante marine, le alghe e i microorganismi che condividono un antenato comune e una serie di caratteristiche morfologiche e fisiologiche simili.

Per esemplificare questa idea, possiamo fare riferimento ad una classificazione schematica degli organismi marini, basata sul livello di tassonomia: questo permette agli esperti di identificare e comprendere meglio la diversità della vita marina e le relazioni evolutive tra le diverse specie: il dominio, definito come un *taxon*, è quindi – in biologia marina – considerato come un criterio per classificare e raggruppare le specie.

Tuttavia, va sottolineato che la classificazione e la comprensione del dominio in biologia marina sono soggette a continue revisioni e aggiornamenti in base alle nuove scoperte scientifiche e alle ricerche in corso nel campo. Questo sottolinea l'importanza della collaborazione tra esperti del settore e ricercatori, per sviluppare e mantenere aggiornate risorse terminologiche e lessicografiche nel campo della biologia marina[74].

1.3.3. Identificazione dei sottodomini di ricerca: dal "Regno Animale" alla categorizzazione per "Famiglie"

Ogni dominio, con il proprio sistema di concetti e nozioni, può essere considerato una struttura autonoma, ma al tempo stesso può contenere al suo interno quelle che Delavigne definisce « sous-structures », ciascuna riferita ad un « sous-domaine en

[74] In questo contesto, l'importanza di consultare *gli esperti, i professionisti del settore*, come i professori Roberto Sandulli (Università degli studi di Napoli "Parthenope") e Ferdinando Boero (Direttore del Museo "Darwin-Dohrn" di Napoli), è stata cruciale per garantire l'accuratezza e la completezza della classificazione degli organismi marini e per informare la pratica terminologica nella disciplina. Questo approccio interdisciplinare, che integra conoscenze teoriche con esperienze pratiche e collaborazioni con esperti del settore, è fondamentale per avanzare nella comprensione della biologia marina e per sviluppare risorse terminologiche e lessicografiche che riflettano accuratamente la ricca diversità della vita marina.

particulier »[75]. Ad esempio, nel caso della biologia ed in particolare di quella marina – che rappresenta il dominio principale in questo contesto – è possibile individuare un sottodominio dedicato alla fauna marina (*i.e.* "Regno Animale"). Questo a sua volta può essere suddiviso in ulteriori sottodomini, come quello delle specie marine, ed altri (*i.e.* livelli tassonomici), ognuno dei quali rappresenta un livello specifico e più dettagliato di classificazione.

Pertanto, constatate sia le difficoltà di delimitare e classificare un *domaine* negli studi linguistici, sia la copiosità dei Domini in biologia e la ricchezza di specie contenute in ognuno di essi, abbiamo deciso di non considerare un intero Dominio in biologia marina, bensì di prendere in esame solo una parte del Dominio, quello degli "Eucarioti" – a cui appartiene anche l'uomo – contenente il "Regno Animale" (§ Fig. 1.4.), come rappresentazione più rilevante in fede alle tematiche alla base del progetto di tesi: il "Regno Animale" (Fig. 1.5.), infatti comprende al suo interno un numero elevato di specie attenzionate nei programmi di valorizzazione del territorio e del patrimonio naturalistico marino[76].

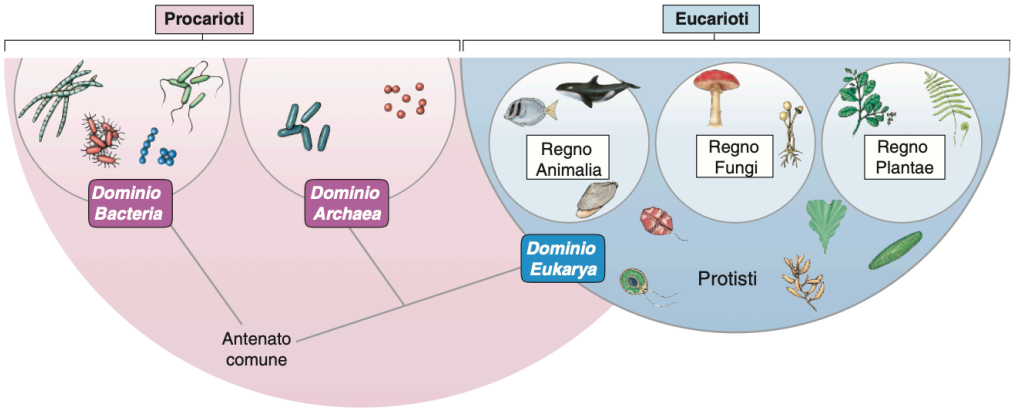


Fig. 1.5. Rappresentazione del più grande raggruppamento di organismi viventi (Sandulli 2011, p. 97).

[75] Delavigne V. 2002, *op. cit.*, p. 7.

[76] Le specie marine e la varietà degli organismi che vivono la cosiddetta "colonna d'acqua", sono indecifrabili, è importante – però – classificarle. Secondo quanto riportato dal Professore Ferdinando Boero durante un incontro presso il museo Darwin Dohrn di Napoli, gli scopi principali della classificazione degli organismi, oltre al nostro (ovvero quello di costituire un corpus quanto più specialistico possibile), sono: dare un nome universale a tutti gli organismi, definire una specie biologica, ovvero con stesse caratteristiche ed interfertilità, e ricostruire una *filoxenia,* ovvero lo studio della filogenesi, la parentela tra gruppi di organismi (con "antenati comuni").

Successivamente, considerando l'aspetto terminologico – secondo quanto illustrato da Buvet *et al.* – come « transdomaine »[77], ovvero come un elemento che specifica un aspetto particolare dell'informazione fornita dall'indicazione di dominio (in questo caso il "Regno Animale"), ed alla luce della vastità di quest'ultimo, abbiamo distinto altre sottocategorie della fauna marina: abbiamo, infatti, identificato ulteriori sottodomini del dominio "Regno Animale", basati sulla classificazione scientifica per "Famiglie"[78]. Abbiamo scelto di focalizzarci sul livello tassonomico della Famiglia come unità di classificazione[79], poiché molti studiosi, tra cui Le Danois, concordano sulla solidità e sulla rappresentatività di questo livello, « les familles […] restent l'élément essentiel du classement méthodique »[80], che offre una panoramica dettagliata delle relazioni filogenetiche tra le specie, consentendo una comprensione più approfondita dell'evoluzione e della diversità all'interno del "Regno Animale" marino, attraverso un focus mirato sulle specie che popolano i nostri mari.

[77] Buvet P.-A., Mathieu-Colas M., « Les champs 'domaine' et 'sous-domaine' dans les dictionnaires électroniques », in *Cahiers de Lexicologie*, vol. 75, Classiques Garnier, Paris 1999, p. 7. Ovvero considerato come un elemento che specifica un aspetto particolare dell'informazione fornita dall'indicazione di dominio (in questo caso il "Regno Animale").

[78] L'ulteriore approfondimento e la successiva ripartizione del dominio "Regno Animale" nei sottodomini "Famiglie", ci ha permesso di effettuare una ricerca molto più mirata e di raccogliere materiale testuale su ogni specie marina.

[79] In biologia, così come in zoologia, in botanica, in protistologia, in batteriologia e in virologia, nel quadro della classificazione tassonomica, la *Famiglia* è uno dei livelli di classificazione scientifica degli organismi viventi e di altre entità biologiche. Più "Famiglie" possono far parte di un unico Ordine. (Wikipedia, Famiglia (tassonomia), 2023).

[80] Le Danois E., « Océanographie, biologie marine et pêches. Remarques ichthyologiques », *Revue des Travaux de l'Institut des Pêches Maritimes*, 13, 1-4, 1939; p. 57.

Livello tassonomico	Esempio
Uomo	
Dominio	Eukaria
Regno	Animalia
Phylum	Chordata
Classe	Mammalia
Ordine	Primates
Famiglia	Hominidae
Genere	*Homo*
Specie e sottospecie	*Homo sapiens sapiens*
Tursiope	
Dominio	Eukaria
Regno	Animalia
Phylum	Chordata
Classe	Mammalia
Ordine	Cetacea
Famiglia	Delphinidae
Genere	*Tursiops*
Specie e sottospecie	*Tursiops truncatus*

Tursiops truncatus, il tursiope

Fig. 1.6. Esempio di classificazione di un uomo e di un delfino, basato sul livello tassonomico della Biologia Marina (Sandulli 2011, p. 97).

Nello specifico, sono state identificate le seguenti "Famiglie" di organismi marini viventi, che per maggiore chiarezza riportiamo di seguito con l'indicazione della specie più rappresentativa di ciascuna:

1. **TORTUES MARINES** (appartenenti alla famiglia Chelonidi);
2. **CORAUX** (appartenenti alla famiglia Coralliidi);
3. **ÉTOILES ROUGES MARINES** (appartenenti alla famiglia Echinasteridi);
4. **DAUPHINS** (appartenenti alla famiglia Delfinidi);
5. **PHOQUES DE MER** (appartenenti alla famiglia Focidi).

Ne consegue che la nostra classificazione può essere letta secondo due prospettive differenti (§ Fig. 1.7.):

I. prospettiva extra-linguistica (esperti del settore): il Dominio Eucarioti viene suddiviso in tre Regni (tra cui il "Regno Animale", § Fig. 1.5) e, successivamente, in "Famiglie", (tra cui le cinque selezionate ai fini del presente studio);

II. prospettiva linguistica (terminologo): il "Regno Animale" è assunto come dominio di riferimento (in quanto il Dominio Eucarioti risulterebbe troppo

vasto per gli obiettivi della ricerca) e le “Famiglie” sono considerate come sottodomini.

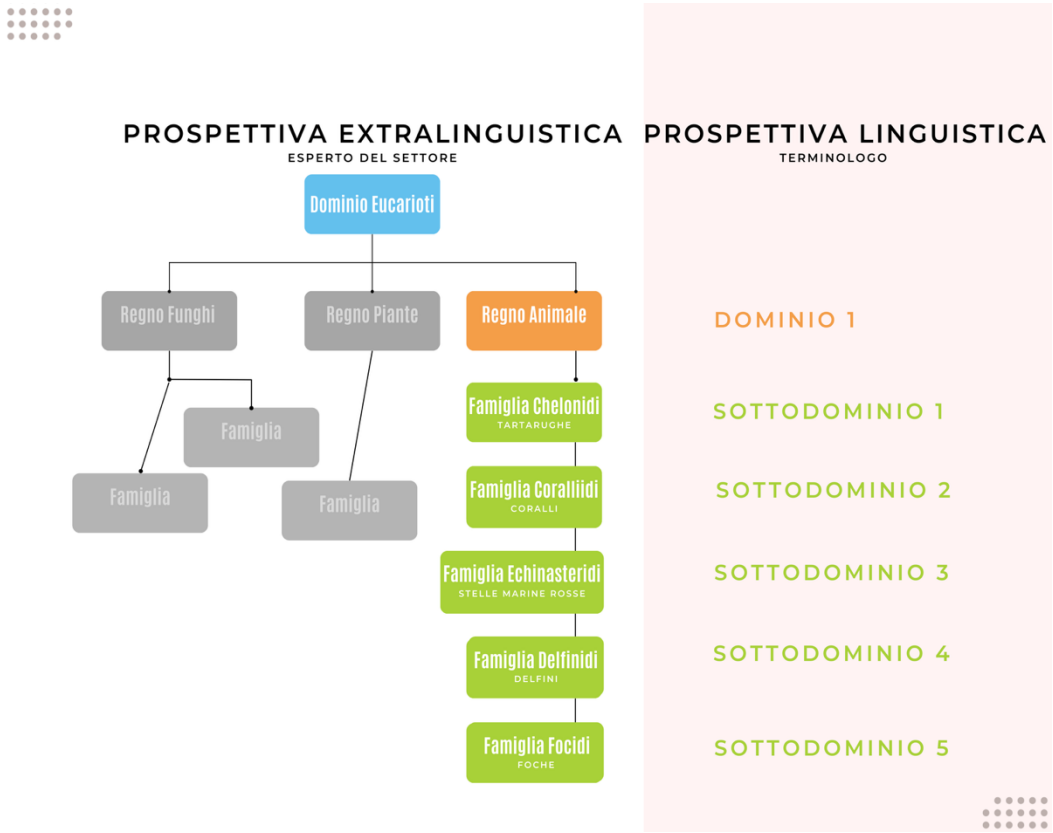


Fig. 1.7. Categorie tassonomiche in biologia marina selezionate per la costituzione del nostro corpus: confronto tra il punto di vista dell’esperto del settore e del terminologo. (Grafica realizzata attraverso Canva).

Alla base di questa impostazione vi è da un lato la volontà di costituire il corpus in maniera graduale, seguendo un criterio di raccolta dei testi fondato su una gerarchia tassonomica (§ paragrafo raccolta testi); dall’altro quella di mantenere un focus terminologico. Lavorare al livello della Famiglia ci ha consentito di cogliere con chiarezza le peculiarità linguistiche legate a ciascun gruppo di organismi, cogliendone tutte le sfumature linguistiche, come ad esempio le nomenclature scientifiche, i luoghi di appartenenza o provenienza, le attività principali svolte, le caratteristiche fisiologiche e tutti gli altri aspetti che possono accrescere il bagaglio

terminologico della stessa specie, mantenendo allo stesso tempo una visione d'insieme della fauna marina[81].

Inoltre, l'inclusione delle "Famiglie" selezionate risponde a tre motivazioni principali: come suggerisce Zollo, a causa della « présence massive de ces espèces dans nos mers », della « leur importance dans le monde de la recherche, de l'éducation et de la science citoyenne » e soprattutto a seguito dei

> récents développements théoriques et pratiques que la biologie de la conservation a connus en Europe les deux dernières décennies grâce aux plans d'actions et aux stratégies politiques de préservation des animaux présents dans les eaux marines.[82]

Infatti, ad esempio, la famiglia Chelonidi comprende le tartarughe marine, specie iconiche che sono spesso considerate simboli della conservazione degli ecosistemi marini. Allo stesso modo, la famiglia Delfinidi, che include i delfini, è di grande interesse scientifico e sociale, poiché queste creature, come anche altre, sono studiate per comprendere le dinamiche comportamentali e sociali degli animali marini[83]. La famiglia Coralliidi invece, fa riferimento ad un prodotto tipico proveniente dalle coste napoletane e che occupa un posto di rilievo nella tradizione, nella cultura e nell'apporto economico della nostra città: il corallo rosso.

Dunque, il nostro approccio alla selezione delle "Famiglie" è stato guidato da una combinazione di considerazioni scientifiche, ecologiche e pratiche, nate soprattutto dalla consultazione con gli esperti del settore.

---

[81] Questo approccio multidisciplinare ci ha permesso di sviluppare un corpus specializzato che riflette la complessità e la diversità del mondo marino, preparandoci a future estensioni e approfondimenti della ricerca: all'interno del progetto di ricerca Ocean Literacy, sono stati costituiti altri corpus al fine di indagare la terminologia specialistica del patrimonio naturalistico marino attraverso diverse prospettive, ad esempio nel corpus *ThalassoCor* sono stati raccolti i testi riguardanti le energie rinnovabili marine, in *EnviDroitCor* sono presenti tutti i testi che argomentano il dibattono sul diritto ambientale o che riportano la normalizzazione della protezione delle specie marine e delle aree marine protette.

[82] Zollo S. D. 2024a, *op. cit*., p. 233.

[83] Halpern B. S. *et al*., « A global map of human impact on marine ecosystems », in *Science*, vol. 319, n. 5865, American Association for the Advancement of Science, New York 2008, pp. 948-952.

CAPITOLO 2. IL CORPUS *ZOOCOR (fr-it)*

Abstract

Il secondo capitolo approfondisce le fondamenta teorico-metodologiche della linguistica dei corpora, delineando le scelte e le fasi operative che hanno condotto alla costituzione del corpus *ZooCor* (fr-it), con una particolare attenzione alla suddivisione tematica in specifiche sezioni per sottodomini.
Successivamente, viene definita l'implementazione del software *CorpusBuilder*, calibrato sulle esigenze del corpus *ZooCor* e sviluppato al fine di automatizzare i principali processi di analisi di corpora.

2.1 Fondamenti teorico-metodologici della linguistica dei corpora

Considerata da molti studiosi, tra cui Rundell & Stock, come una "rivoluzione"[84] all'inizio degli anni Novanta, la linguistica dei corpora non rappresenta una teoria autonoma, bensì un approccio metodologico composto da procedure utilizzabili in combinazione con altri strumenti analitici[85] – che varia in base al fine ultimo dei dati che saranno poi estratti dal corpus stesso[86] – il cui sviluppo è stato reso possibile dall'evoluzione delle tecnologie informatiche e dei sistemi di archiviazione digitale, che ne costituiscono un presupposto imprescindibile[87].

L'Homme definisce un corpus come un insieme di testi rappresentativi di un determinato ambito, scelti dal terminografo per descriverne la terminologia, e – successivamente – stabilisce le caratteristiche alla base della costruzione di esso; infatti, l'autrice riporta che, in primo luogo, un corpus deve costituire « un ensemble de données linguistiques (des mots, des phrases, des morphèmes, etc.) », che devono apparire in un ambiente « naturel (des mots sont combinés à d'autres, sont

[84] Rundell & Stock 1992, in Ferro M. C., «Il corpus RU_SEAH. La lingua russa per la comunicazione specializzata nel settore dell'architettura e delle costruzioni», Educazione Linguistica Language Education, 11, 2, 2022, p. 246.

[85] La varietà metodologica della linguistica dei corpora si manifesta nei diversi approcci all'utilizzo dei dati linguistici, come l'aggiornamento di un *monitor corpus* (Sinclair J., *Corpus, Concordance, Collocation*, Oxford University Press, Oxford, 1991, p. 26; Lenci A. *et al*., *Testo e computer. Introduzione alla linguistica computazionale*. Carocci editore, 2005, p. 33) o la costruzione di un *corpus de référence* (Siepmann D., et al., «Le Corpus de référence du français contemporain (CRFC), un corpus massif du français largement diversifié par genres», in *SHS Web of Conferences*, vol. 27, EDP Sciences, 2016, p. 2; Leech, G., "Principles and applications of Corpus Linguistics", in Viana, V., Zyngier, S., & Barnbrook, G. (Eds.), *Perspectives on corpus linguistics*, Philadelphia, PA, John Benjamins, pp. 155-170, 2011, p. 136; Lopez S., «Corpus de référence et corpus d'usages : méthodologie de constitution pour une analyse des communications pilote-contrôleur», *Cahiers de praxématique*, 54-55, 2010, p. 60).

[86] Si rimanda agli studi sulla ciclicità del lavoro sui corpora (Biber D., "Representativeness in Corpus Design", *Literary and Linguistic Computing*, VIII, 1993, p. 256), alla sottocategorizzazione della stessa metodologia negli approcci *corpus-based* e *corpus driven* (Tognini-Bonelli E., *Corpus Linguistics at Work*, Amsterdam, John Benjamins, 2001, p.65).

[87] Come documentato dagli studi che attribuiscono la nascita della linguistica dei corpora propriamente all'avanzamento dei processi di archiviazione di "some set of machine-readable texts" (McEnery T. & Wilson A., *Corpus Linguistics. An Introduction*, Edinburgh University Press, Edinburgh, 1996, p.1), quindi "in electronic form" (Bowker L., Pearson J., *Working with Specialized Languages. A Practical Guide to Using Corpora*, Routledge, London – New York, 2002, p. 9), all'interno di "computerized databases" (Kennedy G., *An introduction to corpus linguistics*, Routledge, London, 1998, p. 1).

utilisés dans des phrases, les phrases s'agencent dans un texte, etc.) »[88]: attraverso questo punto di vista, L'Homme suggerisce che la selezione dei testi contenenti tali dati linguistici deve basarsi su criteri espliciti, che consentano ad un terzo « d'interpréter les éventuelles généralisations » fatte dal corpus, nel quale tutti i testi sono rappresentativi « de ce qu'on souhaite observer », e che viene dunque assemblato in base all'oggetto in studio[89].

Parimenti, Sinclair specifica che un corpus è una raccolta di testi autentici e leggibili *a macchina*, scelti al fine di "characterize a state or variety of a language"[90]: l'autore mette in luce il ruolo dei corpora come strumenti determinanti poiché permettono di osservare tendenze linguistiche, identificare usi tipici e casi eccezionali, ed ottenere risultati più rapidi ed accurati rispetto ai metodi manuali.

Per questo motivo, un ruolo di particolare rilievo nella costituzione di un corpus è assunto dai criteri di progettazione dello stesso, che abbiamo identificato ed accorpato durante lo studio delle molteplici teorie[91]:

— criterio dimensionale: un corpus deve essere sufficientemente ampio per garantire che le analisi prodotte siano affidabili e statisticamente significative[92];
— criterio di rappresentatività: fa invece riferimento al linguaggio che si intende studiare, sia esso il linguaggio d'uso generale, o un genere specifico[93];

---

[88] L'Homme M.-C., *La terminologie : principes et techniques*, Presses de l'Université de Montréal, Montréal 2004, p. 123.

[89] Inoltre, L'Homme permette di differenziare il corpus dalle opere di consultazione come i dizionari, che, invece, sono il risultato di analisi effettuate da specialisti e riflettono determinate scelte compiute da questi ultimi.

[90] Sinclair J., *op.cit*, p.171.

[91] Nonostante concetti come equilibrio, rappresentatività e comparabilità siano spesso considerati ideali a cui aspirare nella costruzione dei corpora, è raro — se non impossibile — che vengano raggiunti pienamente, definendo tali caratteristiche non come come valori assoluti ma come dimensioni scalari (si rimanda agli studi di Váradi 2001; Biber 1993; Leech 2007, 2011; McEnery & Wilson 2001).

[92] Per quanto concerne i criteri di soglia minima (circa 30-40.0000 *word tokens*) si rinvia agli studi di L'Homme 2004, Sinclair 1991, *et al*. Ciononostante, autori come Bowker e Pearson sostengono che "there are no hard and fast rules that can be followed to determine the ideal size of a corpus" (2002, p.45), ma ammettono la necessità di stabilire una "Corpus size", tra i 10.000 ed alcune centinaia di migliaia di parole.

[93] Tra gli studi che abbiamo preso in considerazione, emerge l'approccio con taglio chiaro ed accessibile – volto a fornire strumenti di base a studiosi della lingua non necessariamente esperti – di Barbera, il quale in particolare sottolinea che il criterio della rappresentatività può essere operazionalizzato tramite un'attenta selezione di "un sample, della lingua analizzata che ne riproduca idealmente, seppur "in miniatura", tutte le caratteristiche, pur nell'impossibilità di avere,

— criterio di annotazione: informazioni sui singoli testi (come i metadati, *i.e.* fonte dei testi, autori, lingue, date, etc.) (§ 2.2.3) che rendono il corpus non solo una raccolta di testi, ma anche una risorsa analitica estremamente versatile, permettendo di effettuare ricerche sia qualitative che quantitative, anche grazie all'uso di strumenti informatici;
— criterio di formattazione: riguarda la leggibilità a macchina (computer), per garantire che i dati possano essere processati da software linguistici, come programmi di concordanze o strumenti di analisi grammaticale: questo implica che i testi devono essere codificati in un formato standard (*i.e.* TXT, XML), privo di errori e compatibile con gli strumenti utilizzati (§ 2.2.2)[94].

Il nostro studio, in particolare, si focalizza sull'uso dei corpora specialistici per il linguaggio con scopo specifico, quello che viene definito da numerosi studiosi come LS, acronimo di *langue de spécialité*: l'insieme delle teorie che hanno pervaso l'ultimo ventennio (tra cui L'Homme[95], Humbley[96], Jacobi[97], etc.) ha prodotto un modello che, concentrandosi su un particolare aspetto della lingua, adopera un linguaggio utilizzato per discutere di ambiti della conoscenza specialistici, e la sua applicazione si dirotta verso la costruzione – ad esempio – di glossari, tesauri, vocabolari, oppure ai fini di un'estrazione terminologica. Inoltre, esso predispone l'utilizzo dei corpora come, da un lato, supporto per la scrittura e la traduzione, e dall'altro come strumenti personalizzabili per l'analisi linguistica.

Di conseguenza, tali corpora rappresentano – come suggerito da L'Homme – « la collecte d'une documentation représentative du domaine »[98] che si vuole descrivere ed il suo utilizzo sono i primi passi di una vera e propria ricerca: quindi un corpus costituisce l'insieme di « *textes représentatifs* du domaine dont [le terminologue]

in ultima analisi, le stesse uguali ed identiche caratteristiche della lingua oggetto di analisi". Barbera M., «Linguistica dei corpora», in Iannaccaro I. (a cura di), 2013, p.44.

[94] Nel suo studio, Viganò analizza dettagliatamente il formato elettronico dei corpora, mettendone in luce i principali vantaggi: "tra cui, innanzitutto, la possibilità di consultarli per mezzo di strumenti informatici che velocizzano il processo di ricerca dell'oggetto di studio e restituiscono dati quantitativi in tempi brevi". Viganò, P.B., "I corpora e il loro sfruttamento in didattica", *Italiano LinguaDue*, vol.2, 2011, p. 116.

[95] L'Homme M.-C. 2004, *op. cit*.

[96] Humbley J., « Le terminologue et le spécialiste de domaine », *ASp. La revue du GERAS*, 1998.

[97] Jacobi D., *La communication scientifique. Discours, figures, modèles*, Presses Universitaires de Grenoble, Grenoble 1999.

[98] L'Homme M.-C. 2004, *op. cit*., p. 119.

compte décrire la terminologie » ed in cui la selezione stessa dei testi deve « reposer sur des critères explicites »[99].

L'autrice, in particolare, sottolinea quanto un corpus costruito sulla base di testi redatti attraverso l'uso di un linguaggio specialistico, che da ora in poi chiameremo « corpus spécialisé »[100], differisca da un corpus caratterizzato da un linguaggio generico, usato nel quotidiano per descrivere argomenti comuni in situazioni pressoché ordinarie, « corpus général », non solo per il vocabolario impiegato, ma anche per le peculiarità stilistiche e le collocazioni che caratterizzano determinate tipologie di testi, contenuti o lingue settoriali:

> en plus de fournir les termes eux-mêmes, les textes spécialisés renferment d'autres données terminologiques qui serviront à mieux saisir leur sens à caractériser leur comportement.[101]

L'accesso ad un *corpus spécialisé*, quindi, consente di individuare rapidamente le strutture sintattiche che distinguono un genere o un tipo testuale specifico, facilitando l'identificazione di tratti distintivi che confermano intuizioni linguistiche sugli usi specifici della lingua. Di conseguenza, attraverso la costituzione di un corpus specialistico è possibile individuare il comportamento combinatorio delle parole: le collocazioni rappresentano un ruolo di particolare rilevanza nelle *langues spécialisées*, considerando la frequente tendenza dei termini di uso specialistico ad essere accompagnati da un secondo termine. Al tempo stesso, infatti – ai fini dell'estrazione terminologica – l'uso dei corpora specialistici può facilitare proprio l'identificazione di termini specialistici. Oltretutto, L'Homme spiega la particolare rilevanza – per la maggior parte dei professionisti coinvolti nell'insegnamento o nell'apprendimento di un qualche tipo di lingua specifica – in modo particolare dell'estrazione dei termini: confrontando il corpus composto da *textes spécialisés* con un corpus di riferimento generale, si possono facilmente rilevare parole con una frequenza superiore al solito che potrebbero essere termini specializzati, non familiari:

---

[99] *Ivi*, p. 123.
[100] *Ivi*, p. 122.
[101] *Ivi*, p. 120.

> Les textes spécialisés fournissent des attestations des termes, c'est-à-dire une preuve qu'ils existent et qu'ils sont effectivement utilisés par les spécialistes. De plus, ils informent sur la fréquence d'emploi d'un terme.[102]

Infatti, ciò viene reso possibile attraverso l'elaborazione di elenchi di parole estratte dal corpus, con l'indicazione della relativa frequenza con la quale appaiono (nel nostro studio l'estrazione automatica è stata realizzata da un software § capitolo 4). I corpora ed i corpora specialistici, dunque, permettono non solo di comprendere il significato dei termini, ma anche di analizzare il loro utilizzo in modo dinamico, con dati sempre aggiornati (relativamente al periodo di riferimento di raccolta dei testi del corpus, c.d. "criterio temporale" § 2.2.2), facilitando soprattutto la comunicazione tra interlocutori di differenti livelli di esperienza di un dato campo di specialità[103].

È proprio in questa prospettiva che si inserisce il nostro studio focalizzato soprattutto sulla divulgazione della conoscenza e sul dialogo tra figure appartenenti a diverse sfere della società (scientifica, politica, istituzionale e mediatica) – come leader politici, esperti, divulgatori scientifici e cittadini – quale chiave per l'avanzamento del progresso e della valorizzazione del patrimonio naturalistico marino dei nostri territori.

Ne consegue che, se come riportato da Humbley « la terminologie constitue, de ce point de vue, le pont entre la profession et la pratique des langues »[104], la realizzazione di un progetto di alfabetizzazione oceanica[105] – attraverso la creazione di risorse linguistiche ed ontologiche tramite le tecnologie digitali – punta alla realizzazione di un primo livello di diffusione e valorizzazione del patrimonio fragile e minacciato della biodiversità marina.

---

[102] *Ivi*, p. 121.

[103] Di conseguenza, possiamo identificare tre tipologie di utenti: gli esperti del settore, ovvero coloro che possiedono una formazione o un'esperienza avanzata nel dominio in questione (da non confondere con i professiosti e/o i "tecnici", Cortelazzo M., *Lingue speciali: la dimensione verticale*, Unipress, Padova, 1994, p. 20); gli studenti, gli esperti di settori correlati ("semi-experts", Bowker e Pearson 2002, *op. cit*., p. 27); i non esperti, tutti coloro che non hanno alcuna familiarità con la *langue de spécialité* in questione.

[104] Humbley J. 1998, *op. cit*., p. 1.

[105] "Ocean Literacy", Zollo S. D. 2024a, *op. cit*.

2.2 Fasi di costituzione e trattamento del corpus *ZooCor*

Il corpus *ZooCor* è stato creato in base alle esigenze linguistiche e di apprendimento dei futuri discenti ed utenti delle *open education resources* (§ 2.1.1.), seguendo una metodologia che Cortellazzo pone su due assi principali:

> […] un'articolazione orizzontale, differenziando l'analisi in relazione alla varietà dei contenuti (quindi lingua della fisica vs lingua della chimica vs lingua dell'economia ...) e procedendo anche all'individuazione di sotto-settori […]. Poi la differenziazione si è estesa in direzione della stratificazione verticale, sociolinguistica, secondo modelli sempre più elaborati […] dovuta principalmente all'attenzione riservata a due forme di uso sociale di tali lingue, e precisamente la divulgazione e l'insegnamento (anche se non va dimenticato che, fin dall'inizio, un peso di rilievo ha la comunicazione tra "tecnico" e consumatore).[106]

Infatti, il corpus si presenta bilanciato sia orizzontalmente, ovvero copre una serie di argomenti diversi relativi a diversi ecosistemi marini nel campo della biologia marina, sia verticalmente, ovvero include materiali di diversi registri, intesi nella più recente accezione della terminologia di Sanga come "livelli di realizzazione e di integrazione [dei sistemi linguistici] dipendenti dalla capacità linguistica del parlante, dall'ambiente e dalle necessità sociali e dai bisogni espressivi"[107], dunque un corpus costruito anche a seconda del grado di specializzazione dei partecipanti alla comunicazione, da quelli prodotti da accademici e professionisti a quelli rivolti a colleghi e studenti, includendo inoltre quelli finalizzati alla diffusione della conoscenza, anche al grande pubblico.

In questo paragrafo ci proponiamo di ripercorrere gli *steps* che hanno portato alla realizzazione del corpus *ZooCor*, in particolare mettiamo in luce le tre fasi che abbiamo delineato e seguito ai fini della sua costituzione[108]:

I. Elaborazione di un modello per la costruzione e la gestione del corpus;

[106] Cortelazzo 1994, *op. cit*, pp. 3-4.
[107] Sanga (1984) in Berruto G., *Sociolinguistica dell'italiano contemporaneo*, Carocci, Roma, 1987, p. 111.
[108] Zollo S. D. 2024a, *op. cit*., pp. 10-29.

II. Selezione ed il trattamento testuale;
III. Stoccaggio dei testi e l'etichettatura dei metadati.

2.2.1 Fase 1: Analisi dei bisogni e progettazione di un modello per la costruzione e la gestione del corpus

Durante questa prima fase, abbiamo effettuato un'analisi delle esigenze linguistiche ed extralinguistiche dei possibili futuri fruitori del nostro corpus, mantenendo fede agli obiettivi di ricerca del progetto.

A monte della progettazione di una risorsa terminologica è infatti necessario prevedere una serie di fasi di "identificazione degli obiettivi generali e specifici, degli utenti" e "del tipo di informazioni da rendere disponibili per gli utenti" [109].

In particolare nel nostro studio, la conduzione di una *need analysis* – sviluppata nella fase di progettazione grazie al confronto con professionisti del settore, tra cui gli esperti, i professori ed i ricercatori in biologia marina del Dipartimento di Scienze e Tecnologie (DiST) dell'Università degli studi di Napoli "Parthenope" e della Stazione Zoologica Anton Dohrn di Napoli – ha permesso una prima mappatura dei bisogni linguistici ed extra-linguistici dei futuri fruitori delle risorse in corso di costituzione (*i.e.* cittadini, viaggiatori, operatori e guide turistiche e ambientali), nonché l'identificazione dei domini e dei sottodomini di riferimento (§ 1.3.2, 1.3.3) per affinare la ricerca dei testi e di conseguenza la costituzione del corpus: i risultati raccolti dall'analisi dei bisogni hanno permesso di sviluppare l'architettura generale del corpus, basata su una categorizzazione multilivello[110] che tiene conto dei sottodomini, della lingua, della finestra temporale, del canale di ricezione, del genere testuale, e di altri metadati definiti in un file Excel (§ 2.2.3), al fine di costruire uno strumento efficace per la *target audience*[111].

---

[109] Per ulteriori approfondimenti, si fa riferimento agli studi sui "requisiti fondamentali da valutare" per la costituzione di un glossario sulle energie solari, sviluppato dall'ITC - CNR (Istituto per le Tecnologie della Costruzione del Consiglio Nazionale delle Ricerche) (Artese M. T., Gagliardi I., «Il glossario delle tecnologie solari: una proposta di sviluppo collaborativo», in Zanola M. T. (a cura di), *Costruire un glossario*, Vita e Pensiero, Milano, 2012, pp. 40-42).

[110] Zollo S. D. 2024a, *op. cit*., pp. 10-29.

[111] Adottata anche da Ferro (2022, p. 260), l'espressione "target audience" si riferisce all'identificazione di un pubblico di riferimento a cui è destinato un prodotto, servizio o contenuto, definito in base a caratteristiche come età, interessi e comportamenti. Dall'insieme dei termini semplici *target*, "Termine (che significa propr. «bersaglio, obiettivo») largamente usato nel

Come già riportato in precedenza, Zollo nel suo lavoro intuisce che, negli ultimi anni, gli esperti sono soliti manifestare una consapevolezza sempre più crescente circa la necessità di adottare un approccio interdisciplinare, piuttosto che « isolationniste »[112], nell'elaborazione terminologica soprattutto relativa alla fauna marina. In particolare, il focus è rivolto verso le strategie di educazione e divulgazione, ai fini di una continua trasmissione del sapere specialistico verso il grande pubblico. Nella fattispecie in questione, tale obiettivo è messo a dura prova da una sfida significativa per quanto concerne le interazioni tra lessico specialistico del dominio in questione (fauna marina), e la lingua generale: molti esperti del settore, tra cui Tillier, indentificano tale complessità anche nella modalità con la quale vengono analizzati i legami tra nomi, concetti, termini e categorie tassonomiche:

> On ne peut pas nier que, dans la mesure où la biologie et la société toute entière utilisent le monde vivant en manipulant les concepts taxonomiques désignés par leur nom scientifique, cette stabilité des noms est du plus grand intérêt pour une majorité d'utilisateurs dont les objectifs sont avant tout opérationnels et limités à leur domaine d'activité.[113]

Di conseguenza, l'analisi dei bisogni del settore ha messo in luce un'ulteriore serie di lacune nella divulgazione delle conoscenze, di matrice soprattutto linguistica, quali scarsa standardizzazione del lessico specialistico, difficile accesso ad un pubblico di non esperti, mancata prospettiva multilingue, oppure prevalenza della lingua latina, che prevedono – come soluzione più efficiente ed efficace – la realizzazione di risorse linguistiche multilingue: è in tale contesto che si inserisce *ZooCor*, il corpus di nostra costituzione.

---

linguaggio commerciale, e spec. della pubblicità e del marketing, con le seguenti accezioni: obiettivo che un'azienda si propone di raggiungere" e *audience*, "Il pubblico di un determinato medium", "l'insieme delle persone raggiunte dal messaggio trasmesso, mediante un mezzo di comunicazione di massa" (Treccani consultato nel 2025).

[112] Zollo S. D. 2024a, *op. cit*., p. 232.

[113] Tillier S., «Terminologie et nomenclatures scientifiques : l'exemple de la taxonomie zoologique», *Langages*, 157, 2005, p. 112.

Quest'ultimo è stato strutturato attraverso l'utilizzo dei principali parametri per la classificazione dei corpora più comunemente esposti nei diversi studi, che abbiamo raccolto nelle seguenti categorie, seguendo uno schema riproposto da Lenci *et al.*: generalità, modalità, cronologia, lingua, integrità dei testi, codifica digitale dei testi[114].

Il criterio di generalità (« niveau de spécialisation »[115], per L'Homme) corrisponde all'identificazione di un corpus in un linguaggio specialistico o generale (rispettivamente definiti come livelli *verticale* e *orizzontale*[116]): il corpus *ZooCor* è un corpus specialistico dedicato ad una particolare area del linguaggio tecnico-scientifico nell'ambito della biologia marina. Tale scelta risponde alla necessità di analizzare terminologie specifiche e concetti relativi agli ecosistemi marini, ai loro abitanti e alle dinamiche biologiche. Si concentra dunque su discorsi tecnici, scientifici e divulgativi, garantendo una copertura completa del linguaggio specialistico utilizzato sia in ambiti accademici e scientifici, che non.

Successivamente, il criterio di modalità (« type de document »[117], per L'Homme) sottende la scelta, da parte di chi costruisce il corpus, della tipologia di dati da raccogliere, dunque se tramite testi scritti, trascrizioni di materiale orale. Il nostro corpus raccoglie testi originariamente prodotti in forma scritta. Questa selezione include una varietà di fonti, come libri specialistici, articoli accademici, documenti istituzionali, pubblicazioni di divulgazione scientifica e (talvolta) articoli di giornale: tale eterogeneità garantisce un'ampia rappresentazione dei diversi stili e livelli di formalità che caratterizzano i testi sulla fauna marina. Inoltre, la scelta di costituire un corpus di lingua scritta si allinea perfettamente con l'obiettivo di esaminare le forme stabili e canoniche della terminologia marittima-scientifica.

Per quanto concerne il criterio cronologico (« date de parution »[118], per L'Homme), dunque sulla raccolta sincronica o diacronica dei testi, il corpus *ZooCor* è stato progettato come corpus diacronico, dunque, includendo testi scritti in epoche

[114] In particolare, tali categorie sono state definite in uno studio di Lenci (2005). Mentre altri studi, ad esempio, Browker e Pearson (2002) rendono la classificazione più generica (*i.e.* General reference corpus vs special purpose/written vs spoken *etc.*)

[115] L'Homme M.-C. 2004, *op. cit*., p. 126.

[116] Cortellazzo 1994, *op. cit*.

[117] *Ibidem.*

[118] *Ibidem*.

diverse, dal XVIII secolo fino al XXI secolo. Questa scelta permette di tracciare il cambiamento e l'evoluzione della terminologia e dei concetti biologici attraverso i secoli, osservando l'introduzione di nuovi termini, l'obsolescenza di altri e l'evoluzione delle definizioni e della nomenclatura tassonomica. Infine, la strutturazione "temporale" del corpus consente quindi sia analisi linguistiche microtemporali che macrotemporali (§ 2.2.2).

Un'ulteriore distinzione tra le differenti tipologie di corpora è quella che riguarda la lingua (criterio di « langues/ langue de rédaction »[119], per L'Homme). Il nostro corpus *ZooCor* è strutturato come corpus bilingue comparato – contenente quindi testi originali selezionati in francese ed in italiano, seguendo quanto riportato da Lenci:

> Un corpus bilingue (multilingue) *comparabile* non contiene invece testi in traduzione, ma testi originali in lingue diverse […] Il corpus è comparabile nella misura in cui i criteri di selezione dei testi sono gli stessi nelle varie lingue[120].

Per quanto riguarda l'integrità dei testi ("Text extracts vs full texts"[121], in Bowker e Pearson), il corpus contiene testi interi, evitando l'uso di frammenti che potrebbero ridurre la rappresentatività ed alterare la coerenza tematica; infatti la scelta di includere testi completi assicura una visione più accurata della struttura linguistica e dei registri adottati, conservando contesti essenziali per l'analisi dei termini e per l'interpretazione dei dati testuali: "In LSP studies, however, the concepts, terms, patterns and contexts that interest you might appear in any section of a text. In fact, their location in a text may be highly relevant"[122].

Infine, *ZooCor* si presenta come un corpus codificato, ovvero sottoposto ad una codifica digitale dei testi: ogni testo è etichettato « selon un ensemble de […] classes de métadonnées textuelles »[123] raccolti in un file Excel (§ 2.2.3), che

[119] *Ibidem.*
[120] Lenci A. et al., *Testo e computer. Introduzione alla linguistica computazionale*. Carocci editore, 2005, *op. cit*., p.31.
[121] Bowker L. & Pearson J. 2002, *op. cit*., p.49.
[122] *Ibidem*.
[123] Zollo S. D. 2024a, *op. cit*., p. 234.

includono informazioni essenziali come l'autore, l'anno di pubblicazione, la lingua, il genere testuale e altre caratteristiche pertinenti.

Di seguito, presentiamo una breve sintesi schematica dei criteri di classificazione dei corpora, adottati e modellati per il nostro studio:

Parametri per la classificazione del corpus *ZooCor*	
1. Generalità	CORPUS SPECIALISTICO (VERTICALE) Descrizione di un particolare linguaggio di specialità: discorsi tecnico scientifici sulla biologia marina
2. Modalità	CORPUS DI LINGUA SCRITTA Testi originariamente prodotti in forma scritta: libri, articoli di giornale, testi istituzionali, testi accademici, ma anche in testi di divulgazione (*criterio testuale* § 2.2.2)
3. Cronologia – fattore tempo	CORPUS DIACRONICO Testi appartenenti a periodi storici diversi, con lo scopo di monitorare il cambiamento linguistico su scala microtemporale e/o macrotemporale: dal XVIII al XXI secolo (*criterio temporale* § 2.2.2)
4. Lingua	CORPUS BILINGUE (COMPARATO) Contiene testi originali in francese ed italiano
5. Integrità dei testi	CORPUS INTERO Contiene interi testi
6. Codifica digitale dei testi	CORPUS CODIFICATO Attraverso la raccolta di metadati di ogni testo in un Excel (§ 2.2.3)

Fig. 2.1. Schema di classificazione e struttura del Corpus *ZooCor.*

2.2.2 Fase 2: la selezione ed il trattamento dei testi

Il corpus *ZooCor* è un corpus specializzato comparabile bilingue, che attualmente implica due lingue romanze, il francese e l'italiano, e che riunisce 410 testi scritti, la maggior parte dei quali in lingua francese.

La scelta del francese come lingua *pivot* risponde ad esigenze tecnico-scientifiche, come riferito da Zollo, perché il focus dell'intero progetto verte sulla descrizione della terminologia francese e sulle ricerche in contrastiva con la lingua italiana, da cui emergono differenze significative:

> […] tout le projet est centré, en particulier, sur la description de la terminologie en langue française du domaine en question et sur des recherches contrastives du français par rapport à d'autres langues, parmi lesquelles l'italien. Malgré l'apparente similitude entre les deux systèmes linguistiques, les deux communautés discursives concernées conceptualisent de manière distincte le lexique et les terminologies scientifiques de la faune marine, en raison de leur diversité culturelle en matière de communication et de gestion[124].

In particolare, i testi sono stati raccolti sulla base di tre criteri extralinguistici, ovvero un criterio cronologico, un criterio tematico e un criterio testuale.
Per quanto riguarda il criterio cronologico, abbiamo selezionando testi appartenenti a una finestra temporale ampia, dal XVIII al XXI secolo, al fine di seguire i cambiamenti linguistici ed extralinguistici che si sono verificati nel corso dei secoli in una scala micro e/o macro-temporale di questo dominio[125].

[124] Zollo S. D., 2024a, *op.cit*., p. 233.
[125] Attualmente, il corpus e la raccolta di metadati di ciascun testo permettono di effettuare uno studio di tipo diacronico, tuttavia per scelte metodologiche nel presente elaborato il focus verte maggiormente sulla modellizzazione dei termini da un linguaggio specialistico.

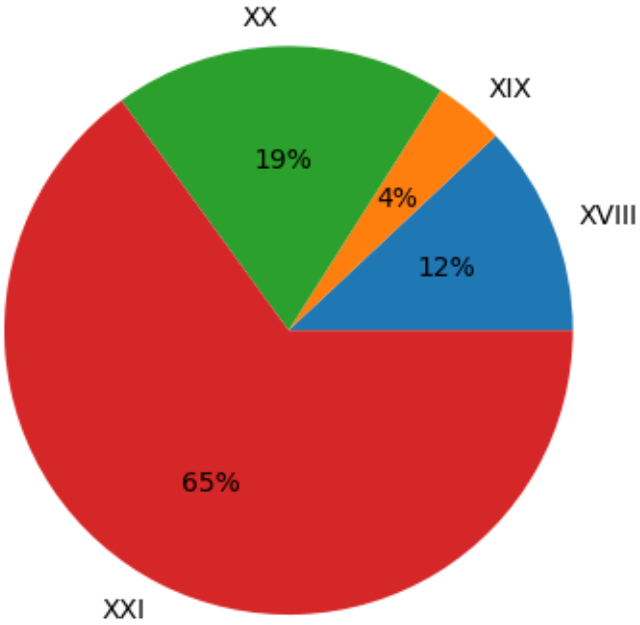


Fig. 2.2. Rappresentazione quantitativa della finestra temporale del corpus *ZooCor* (criterio temporale).

In riferimento al criterio tematico, i testi sono stati selezionati seguendo la suddivisione ed il raggruppamento per sottodomini, ovvero in "Famiglie" di specie e sottospecie marine (§ 1.3.3., § Fig. 2.3.): così facendo, abbiamo cercato di conferire alla raccolta dei testi una rigorosità scientifica, in quanto l'impostazione rispecchia la classificazione tassonomica riconosciuta nelle scienze biologiche, che procede dal generale (Dominio) al particolare (Famiglia, Genere, Specie, sottospecie).

Infatti, per restringere il campo di indagine e la successiva selezione dei documenti[126], abbiamo avviato la ricerca negli archivi digitali attraverso le *mots-clés* generalmente corrispondenti al nome della Famiglia di riferimento (*i.e.* Chelonidi, Coralliidi, Focidi), o della specie di punta (*i.e.* tartarughe marine, coralli rossi, foche). Di conseguenza i testi venivano raccolti per sezioni basate sui sottodomini e successivamente uniti in un unico corpus.

[126] Sarebbe stato infattibile realizzare una ricerca basandosi su termini generici come "biologia marina", "fauna marina", "patrimonio naturalistico marino": in aggiunta, l'esclusività rivolta nei confronti delle specie marine animali è motivata anche dall'attualità degli sviluppi in questo campo di indagine, a causa dei cambiamenti climatici che portano ad una crescente necessità di nuove nomenclature per le specie derivanti da nuovi flussi migratori.

La costituzione progressiva del corpus ha garantito una coerenza interna ed una maggiore rappresentatività rispetto al patrimonio naturalistico marino, rispondendo alle esigenze sia scientifiche che linguistiche del progetto di ricerca (§ 1.3.)[127]: infatti la suddivisione in sezioni ci ha permesso di effettuare alcuni studi pilota per avere contezza dell'apparato terminologico (sia in termini di quantità, che di qualità del lessico estratto), e per verificare il lavoro di analisi terminologica e lessicografica prima su scala minore – tramite l'analisi dei corpora ridotti – e successivamente adattando il lavoro al corpus finale, al fine di avere un primo riscontro della compatibilità con la metodologia da applicare e dell'uso degli strumenti da implementare (§ 5.1.).

Les êtres vivants marins (zoologie marine/faune marine)				
Domaine: Les eucaryiotes (Eukaryota)				
Règne: animal ou Les Animaux (Animalia, Linnaeus, 1758)				
Familles				
CHELONIIDAE (Tortues marines)	CORALLIIDAE (Coraux)	ECHINASTERIDAE (Étoiles rouges marines)	DELPHINIDAE (Dauphins)	PHOCIDAE (Phoques de mer)

Fig. 2.3. Rappresentazione della classificazione dei sottodomini nei metadati di Excel del corpus *ZooCor* (criterio tematico).

Infine, con l'obiettivo di rispondere all'esigenza di rappresentatività e di variabilità, un corpus – come suggerito da Lenci – "[…] deve tenere traccia dell'intero ambito di variabilità dei tratti e proprietà di una lingua. Il modo in cui questi si distribuiscono dipende fortemente dalla tipologia dei testi della lingua"[128].

Per questo motivo, abbiamo strutturato un criterio testuale, la cui esistenza nasce dall'esigenza di recuperare unità terminologiche non solo nel lessico specialistico: l'eterogeneità del corpus si manifesta nella suddivisione della raccolta di testi istituzionali, testi accademici, ma anche in testi di divulgazione.

[127] La limitazione a soli cinque sottodomini si spiega con i vincoli imposti dal progetto di ricerca dottorale e non esclude la possibilità di futuri ampliamenti del corpus, sia con l'inserimento di altre aree della biologia marina, sia con l'estensione di materie a settori disciplinari adiacenti (es. oceanografia, zoologia, ecc.) (vedasi § capitolo 4).

[128] Lenci A. et al. 2005, *op. cit.*, p.36.

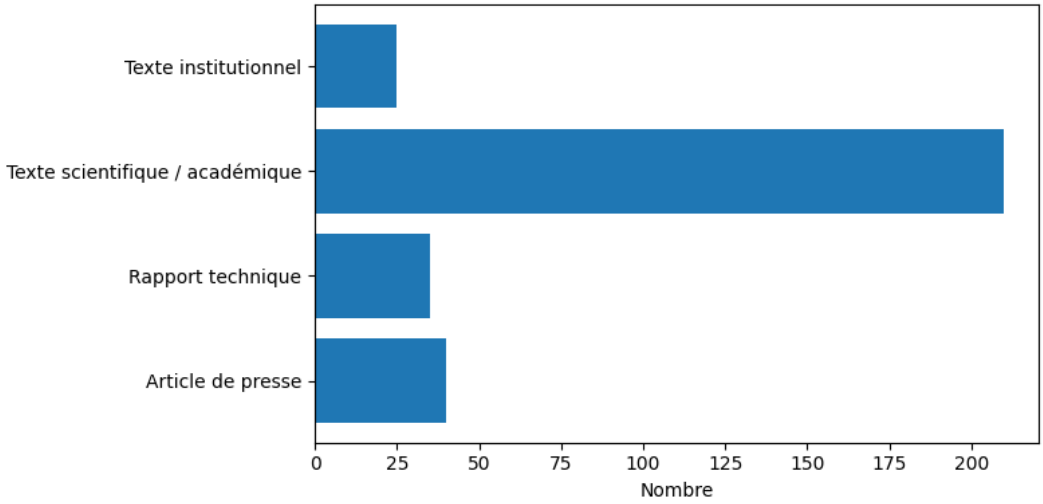


Fig. 2.4. I generi testuali del corpus *ZooCor* (criterio testuale).

La raccolta – con l'obiettivo di essere quanto più rappresentativa del dominio in questione – (Fauna marina, "Regno Animale" nella biologia marina) presenta principalmente testi quali tesi di dottorato, trattati sull'argomento, relazioni di missione, articoli scientifici e saggi, acquisiti in formato elettronico (.pdf), attraverso archivi, biblioteche digitali e banche dati scientifiche: la raccolta ha incluso testi provenienti da diverse istituzioni e *database*, come (citiamo i più frequenti) la Biblioteca Centrale dell'Università Parthenope di Napoli (per i testi in formato cartaceo e poi digitalizzati), la Biblioteca del Mare, ResearchGate, Ifremer, Coral Guardian, HalScience, Gallica, etc. (per i testi in formato digitale).

A titolo d'esempio, di seguito, riportiamo la rappresentazione della distribuzione dei testi basata sulla fonte digitale:

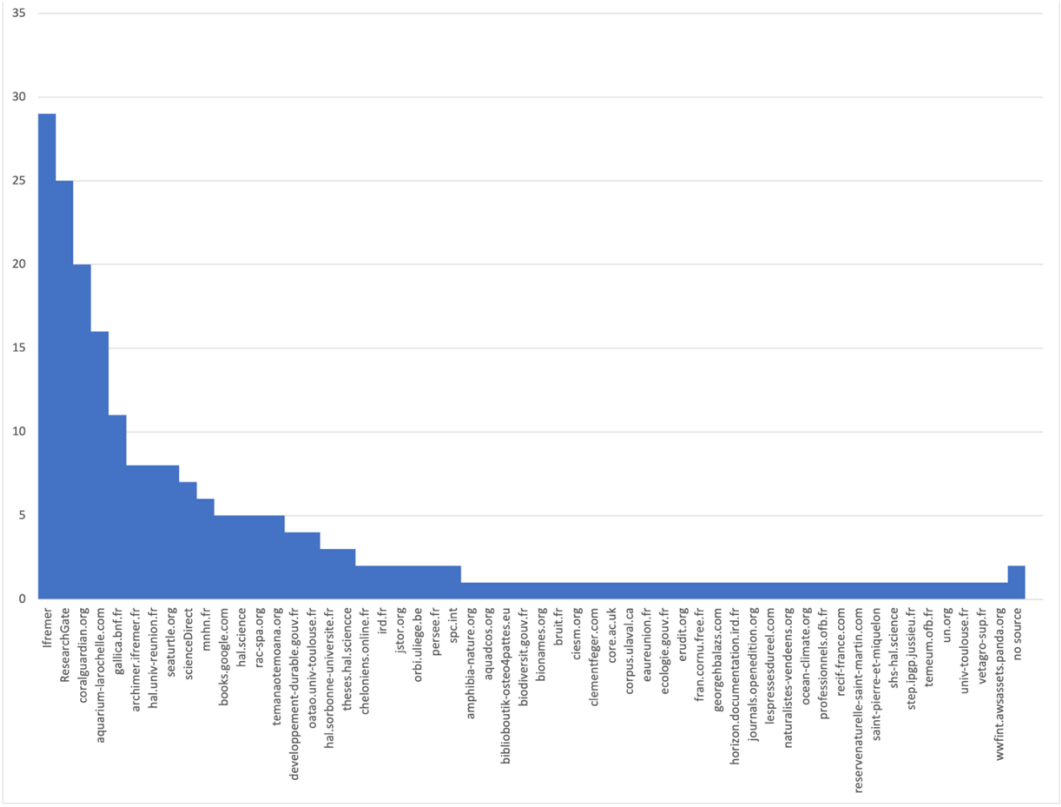


Fig. 2.5. Rappresentazione della distribuzione dei testi basata sulla fonte digitale di provenienza.

Per garantire coerenza e qualità al corpus, ci siamo occupati di escludere tutti i testi raccolti in un primo momento in formato cartaceo, non potendo essere facilmente digitalizzati ed integrati, tutti i testi non pertinenti propriamente all'ambito marino (*i.e.* i testi riguardanti i fiumi, i laghi, ed altri ecosistemi non salini), i testi riguardati prettamente altre specie appartenenti ad altre "Famiglie" del "Regno Animale" che non facessero parte dei cinque sottodomini di ricerca.

La conversione dei testi cartacei in formato elettronico è complessa e dispendiosa; perciò, è sempre preferibile raccogliere materiali già disponibili in formato digitale: il web rappresenta in tal senso una fonte inesauribile di risorse, ma molto complessa da gestire.

La formattazione digitale permette una ricerca veloce, efficiente e affidabile, essenziale per l'analisi di grandi quantità di dati. I testi, raccolti originariamente in formato .pdf, sono stati convertiti successivamente in formato .txt, con codifica

UTF-8, attraverso il software AntFileConverter[129]. Di seguito, ci siamo occupati della "pulizia dei testi": a ciascuno dei testi in formato .txt abbiamo sottratto manualmente gli elementi superflui e non pertinenti come tabelle, note, bibliografie, parole in altre lingue e tutto ciò che ostacolasse dal garantire una migliore qualità dei documenti raccolti. Infine, i testi in entrambi i formati sono stati salvati su un cloud condiviso (OneDrive) e – ai fini di una successiva estrazione terminologica – i testi in formato .txt sono stati uniti tutti in due file: un file .txt contente i documenti in lingua francese ed un file .txt contente il corpus nella sua componente in lingua italiana.

Il corpus *ZooCor* presenta, dunque, 350 documenti in lingua francese, articolati come segue: il 37% di essi si articola attraverso un discorso accademico, il 34% tramite un discorso professionale, ed il 29% con un discorso di divulgazione, per un totale di 425.000 token; analogamente, esso è composto da 110 documenti in lingua italiana, con il 44% dei quali basati su un discorso accademico, il 27% su un discorso professionale, ed il 29% su un discorso di divulgazione, per un totale di 210.000 token[130].

2.2.3 Fase 3: lo stoccaggio dei testi e l'etichettatura dei metadati

Il corpus rappresenta una risorsa estremamente versatile, che permette al suo fruitore di svolgere ricerche sia qualitative che quantitative, attraverso l'ampia fruibilità di dati che deriva da un rispettivo preciso inserimento di materiale informativo in esso.

Infatti, nel nostro studio ci siamo occupati di raccogliere i metadati di ciascun documento del corpus e di strutturarli all'interno di un file Excel, seguendo una classificazione in circa diciassette categorie: dominio/sottodominio di appartenenza, lingua di stesura, titolo del documento, autore, anno di pubblicazione, anno di consultazione, editore, genere testuale (testo istituzionale, testo scientifico/accademico, rapporto di missione, articolo divulgativo), fonte

[129] Anthony L., *AntFileConverter* (Version 2.1.0) [Computer Software]. Tokyo, Japan: Waseda University, (disponibile online: https://www.laurenceanthony.net/software/AntFileConverter), 2024.

[130] Zollo S. D. 2024a, *op. cit*., pp. 10-29.

bibliografica (l'archivio digitale dove è stata effettuata la ricerca), fonte digitale (la fonte specifica da dove è possibile scaricare il documento), formato (.pdf/.txt), numero di pagine, la parola lanciata per la ricerca, il paese di riferimento del testo, il mare in cui è ambientato il soggetto del documento, le parole chiave del testo, eventuali note al documento.

Metadati testuali di ZooCor	Dominio/sottodominio
	Lingua
	Titolo
	Autore
	Anno di pubblicazione
	Anno di consultazione
	Editore
	Genere testuale
	Fonte bibliografica
	Fonte digitale
	Formato
	Numero di pagine
	Parola di ricerca
	Paese
	Mare
	Parole Chiave
	Note

Fig. 2.6. Metadati per la schedatura dei testi del corpus *ZooCor*.

2.3 Strumenti informatici al servizio della linguistica dei corpora: fasi e risultati di una collaborazione interdisciplinare

Parallelamente, nell'ambito del soggiorno di ricerca dottorale presso il laboratorio di informatica dell'Université de Pau et des Pays de l'Adour (LIUPPA), al fine di affrontare sfide metodologiche specifiche per lo sviluppo del presente lavoro di ricerca – in particolare legate alla raccolta, organizzazione e gestione dei testi che compongono un corpus – è stata concretizzata una proficua collaborazione tra il

gruppo di ricerca in Linguistica Francese dell'Università degli Studi di Napoli "Parthenope" ed il gruppo di ricerca in Informatica del LIUPPA[131].

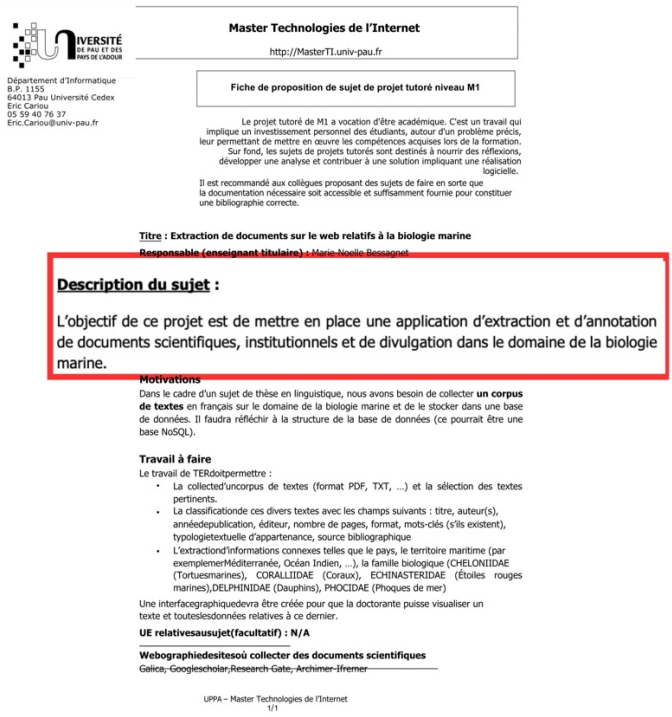
2.3.1 *Fiche de Sujet*: requisiti e criteri per l'implementazione del software *CorpusBuilder*

Sebbene il corpus *ZooCor* fosse già in una fase di costituzione, in prospettiva di studi futuri e contigui si è manifestata l'impellente necessità di plasmare uno strumento che non solo consentisse l'integrazione di nuove risorse testuali relative a sottodomini affini, ma che potesse altresì automatizzare parte dei processi di raccolta e gestione, ampliando così la portata e la rilevanza del corpus per la protezione e la valorizzazione del patrimonio naturalistico marino.

Di conseguenza, l'obiettivo è stato individuato nello sviluppare una risorsa informatica capace di ridurre i tempi di elaborazione manuale e migliorare la qualità e l'accessibilità dei dati terminologici. La sinergia tra i due gruppi di ricerca ha mirato a creare un sistema che potesse costituire un corpus ed organizzare i rispettivi testi in modo dinamico, preparandolo per analisi terminologiche più avanzate e per l'applicazione a scopi divulgativi.

La genesi di questa esigenza si è tradotta nella redazione di una "Fiche de Sujet", un documento programmatico che ha delineato con precisione le esigenze funzionali e non funzionali del software *CorpusBuilder*, partendo – come studio pilota – dall'arricchimento del corpus *ZooCor*.

[131] La complessità crescente della terminologia specialistica e la necessità di gestire grandi volumi di dati testuali rendono indispensabile un approccio multidisciplinare: infatti, come discusso nei primi paragrafi del presente capitolo, la costruzione e l'analisi di corpora specializzati beneficiano enormemente della sinergia tra esperti linguisti e sviluppatori informatici.



Département d'Informatique
B.P. 1155
64013 Pau Université Cedex
Eric Cariou
05 59 40 76 37
Eric.Cariou@univ-pau.fr

Master Technologies de l'Internet

http://MasterTI.univ-pau.fr

Fiche de proposition de sujet de projet tutoré niveau M1

Le projet tutoré de M1 a vocation d'être académique. C'est un travail qui implique un investissement personnel des étudiants, autour d'un problème précis, leur permettant de mettre en œuvre les compétences acquises lors de la formation. Sur fond, les sujets de projets tutorés sont destinés à nourrir des réflexions, développer une analyse et contribuer à une solution impliquant une réalisation logicielle.

Il est recommandé aux collègues proposant des sujets de faire en sorte que la documentation nécessaire soit accessible et suffisamment fournie pour constituer une bibliographie correcte.

Titre : Extraction de documents sur le web relatifs à la biologie marine

Responsable (enseignant titulaire) : Marie-Noelle Bessagnet

Description du sujet :

L'objectif de ce projet est de mettre en place une application d'extraction et d'annotation de documents scientifiques, institutionnels et de divulgation dans le domaine de la biologie marine.

Motivations

Dans le cadre d'un sujet de thèse en linguistique, nous avons besoin de collecter **un corpus de textes** en français sur le domaine de la biologie marine et de le stocker dans une base de données. Il faudra réfléchir à la structure de la base de données (ce pourrait être une base NoSQL).

Travail à faire

Le travail de TERdoitpermettre :

- La collected'uncorpus de textes (format PDF, TXT, ...) et la sélection des textes pertinents.
- La classificationde ces divers textes avec les champs suivants : titre, auteur(s), annéedepublication, éditeur, nombre de pages, format, mots-clés (s'ils existent), typologietextuelle d'appartenance, source bibliographique
- L'extractiond'informations connexes telles que le pays, le territoire maritime (par exemplemerMéditerranée, Océan Indien, ...), la famille biologique (CHELONIIDAE (Tortuesmarines), CORALLIIDAE (Coraux), ECHINASTERIDAE (Étoiles rouges marines),DELPHINIDAE (Dauphins), PHOCIDAE (Phoques de mer)

Une interfacegraphiquedevra être créée pour que la doctorante puisse visualiser un texte et touteslesdonnées relatives à ce dernier.

UE relativesausujet(facultatif) : N/A

Webographiedesitesoù collecter des documents scientifiques

~~Galica, Googlescholar,Research Gate, Archimer-Ifremer~~

UPPA – Master Technologies de l'Internet
1/1

Fig. 2.7. *Fiche de Sujet* redatta dal gruppo di ricerca in informatica per rispondere ai bisogni esposti dall'analisi terminologica.

Tale documento, frutto della collaborazione tra i *co-incadrants* (Marie-Noëlle Bessagnet, Virginia Carrella, Annig Lacayrelle, Maxime Masson, Christian Sallaberry, Silvia Domenica Zollo) e gli studenti del Master 1 Technologies de l'Internet (M1TI) dell'Università di Pau, ha fissato i requisiti per la creazione dell'applicazione dedicata all'automazione di alcune delle fasi di un lavoro di indagine terminologica di domini affini a quello del patrimonio naturalistico marino: le motivazioni sottostanti erano chiaramente connesse al nostro progetto di tesi in linguistica, che, in vista di studi futuri, necessitava di ampliare ed alimentare il corpus *ZooCor* attraverso l'arricchimento di dati su domini specifici e contigui come il "Droit environnemental marin" (con sotto-tematiche quali regolamenti della *pollution acoustique marine, pollution lumineuse, justice climatique, justice environnementale, habitats naturels et semi-naturels, protection des espèces animales, droit de l'environnement marin*) e delle "Espèces marines envahissantes" (presenza, salvaguardia e gestione delle specie marine alloctone e invadenti).

Dunque, sulla base dei criteri di costituzione manuale del corpus *ZooCor* (criterio testuale, criterio temporale, formato dei testi, siti di raccolta dei testi, lingue di

ricerca dei testi) e dei criteri di classificazione ed estrazione stessa in un file Excel dei metadati ricavati dai testi (titolo, autore, anno di pubblicazione, parole chiave, riassunti, fonte bibliografica)[132], al fine di plasmare lo strumento informatico, nella *Fiche de Sujet* è stato definito un *Travail à faire*, nel quale sono stati specificati i precisi steps che gli studenti del M1TI avrebbero dovuto automatizzare, tra cui:

— la raccolta di testi, in formati compatibili come PDF e TXT, con particolare attenzione alla selezione dei testi pertinenti: per fare ciò le fonti di raccolta sono state identificate in webografie specifiche, includendo archivi come HAL Archives, ArchiMer, Research Gate, e biblioteche digitali come Gallica, con una finestra temporale dal 2000 ad oggi. È importante sottolineare come la scelta di fonti digitali risponda alla necessità di superare le complessità e i costi della conversione di testi cartacei, privilegiando materiali già disponibili in formato elettronico per garantire una ricerca veloce, efficiente e affidabile, (paragrafo 2.2.2, "La selezione ed il trattamento dei testi");
— la classificazione dettagliata dei testi: con campi specifici quali titolo, autore/i, anno di pubblicazione (dal 2000 in poi), editore, numero di pagine, formato, parole chiave (se presenti), tipologia testuale (testi giuridici/legislativi, testi accademici, testi istituzionali) e fonte bibliografica. Questi criteri si allineano con quanto esposti in precedenza (paragrafo 2.2.3, "Lo stoccaggio dei testi e l'etichettatura dei metadati") riguardo all'etichettatura dei metadati, essenziale per una successiva interrogazione qualitativa e quantitativa del corpus;
— l'estrazione automatica dei dati testuali.

La *Fiche de Sujet* ha altresì richiesto la creazione di un'interfaccia grafica intuitiva per consentire di visualizzare ogni testo ed i dati ad esso relativi, in entrambe le lingue di riferimento.

2.3.2 Genesi ed architettura del software *CorpusBuilder*

Nell'ambito della collaborazione tra i due atenei e sulla base dei criteri proposti per la costituzione del corpus *ZooCor*, è stato sviluppato *CorpusBuilder*, un'applicazione che permette di automatizzare i processi di raccolta dei testi in

[132] Zollo S. D. 2024a, *op. cit*., pp. 10-29.

lingua francese ed italiana e di estrazione dei metadati dai singoli testi: lo strumento informatico, sviluppato dagli studenti e dai ricercatori del laboratorio di informatica « LIUPPA » dell'UPPA, mira ad agevolare la raccolta, la gestione e l'analisi di dati testuali nel contesto della costituzione di corpora e nello studio di terminologie specialistiche, come nel caso dei sottodomini del patrimonio naturalistico marino.

Fig. 2.8. *Soutenance de mémoire* degli studenti del master presso il LIUPPA che si sono occupati dell'implementazione di *CorpusBuilder* (21.05.2024).

Il software (disponibile sia in lingua italiana che in lingua francese), grazie ad un'interfaccia chiara e funzionale supporta l'utente in tutte le fasi del processo, dalla ricerca di documenti alla gestione e all'esportazione dei dati. L'interfaccia-utente dell'applicazione, infatti, è progettata per garantire un'interazione intuitiva con il sistema e per facilitare l'esecuzione delle varie attività richieste, per cui le funzionalità avanzate di filtraggio e ordinamento, insieme con la possibilità di integrare strumenti esterni come Excel, rendono lo strumento versatile ed efficace per la gestione delle risorse documentali in un dominio specifico.

Per cominciare, l'interfaccia utente dell'applicazione consente un'interazione intuitiva con il sistema, facilitando l'esecuzione delle varie attività richieste, attraverso una prima scelta della lingua con la quale avviare le ricerche[133].

In particolare, l'applicazione è composta da due pagine principali: *Accueil*, che consente di cercare documenti, e *Base de données*, che offre la possibilità di

[133] Infatti, in base alla lingua scelta cambiano anche i comandi, che nel presente capitolo – per praticità – riportiamo maggiormente in lingua francese.

visualizzare il contenuto del database (ovvero dove vengono archiviati i testi selezionati) ed esportare i dati nei vari formati (processo di cui discuteremo più in avanti). Entrambe le pagine hanno un'intestazione simile che ci permette di passare da una pagina all'altra tramite un "click", mentre l'opzione *Quitter* consente di chiudere in sicurezza il server (Fig. 2.9.[134]).

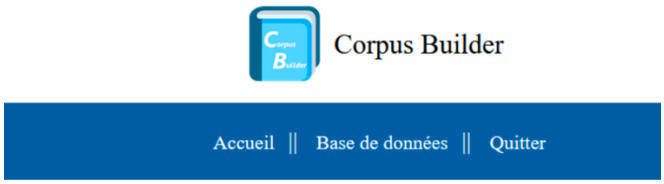


Fig. 2.9. Interfaccia principale con le 3 opzioni di scelta: *Accueil*, *Base de données*, *Quitter*.

La pagina iniziale *Accueil* è dotata di una barra di ricerca centrale accompagnata da tre pulsanti di opzione (« radio-Bouton ») che permettono di scegliere il sito da cui effettuare la ricerca (Hall Science, EurLex, Gallica, WWF, *et al.*): ciò ha consentito di filtrare la ricerca dei testi che, all'interno di un dominio specialistico, siano rilevanti da un punto di vista scientifico, e che dunque costituiscano una fonte primaria di documentazione nella ricerca terminologica[135] (Fig. 2.10.). Quindi, una prima automatizzazione dei processi ha riguardato la ricerca dei testi (che al contrario avveniva manualmente attraverso il consulto di fonti bibliografiche differenti), per cui abbiamo lanciato alcune parole chiave come *pollution marine*, per il francese, ed *inquinamento marino*, per l'italiano.

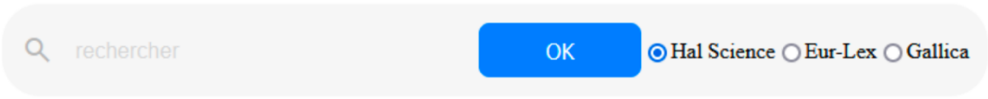


Fig. 2.10. Barra di ricerca implementata graficamente al fine inserire la *mot-clé* e lanciare l'indagine, ma soprattutto per filtrare la ricerca attraverso la selezione tra le fonti.

[134] Alcune delle figure relative alle funzionalità del software di *CorpusBuilder* sono state estratte dal documento con le istruzioni d'uso redatto dagli studenti del M1TI dell'UPPA Alexy Del Amo Alonso, Maxime Silla, (Université De Pau Et Des Pays De L'Adour, Collège Stee, Spécialité : Technologies De L'internet).

[135] Zanola M. T., «Attività terminologica e fonti di documentazione ieri e oggi: problemi e metodi», *mediAzioni*, 16, 2014, disponibile on line: http://mediazioni.sitlec.unibo.it.

Dopo aver eseguito una ricerca, attraverso l'inserimento di una delle parole chiave, il sistema restituisce il numero di documenti trovati ed organizza i risultati con un massimo di 30 documenti per pagina. Ogni risultato è presentato in una scheda composta da vari campi, che richiamano alcuni dei metadati prefissati dallo studio in terminologia ed un comando finale (Fig. 2.11.):

— TITRE : titolo del documento, ordinabile alfabeticamente;
— AUTEUR : autore del documento;
— ANNEE : anno di pubblicazione, ordinabile in ordine crescente o decrescente;
— MOTS CLÉS : parole chiave del documento, che il software estrae automaticamente dai documenti;
— RESUMÉ & LINK : riassunto e collegamento ipertestuale del documento, per questi campi è possibile ordinare la lista dei documenti in base a quelli che sono in possesso, o meno, dei riassunti e/o dei links, ottimizzando la leggibilità;
— LINK : mostra un collegamento cliccabile che apre il documento PDF associato in una nuova scheda del browser;
— AJOUTER : consente di aggiungere il documento al database, ovvero al corpus virtuale (una volta aggiunto, il colore del campo cambia in azzurro, indicando la sua presenza nel database).

Fig. 2.11. Esempio di ricerca di *pollution marine.*

Il comando finale "AJOUTER" dà la possibilità di selezionare i testi ed "aggiungerli" in un database (chiamato *Base de données* nell'interfaccia francese e *Database* nell'interfaccia italiana) (Fig. 2.12.), ovvero una seconda interfaccia del software in cui vengono inseriti i documenti selezionati, ciò consente l'organizzazione e l'archiviazione delle informazioni, proprio come in un corpus virtuale costituito all'interno dell'applicazione, al quale accedere per consultare ed estrarre dati nelle

diverse modalità (senza dover scaricare tutti i file sul proprio dispositivo elettronico) (Fig. 2.13.): questo procedimento è agevolato anche dalla rappresentazione sulla scheda di "MOTS CLÉS" e "RESUMÉ", quindi delle parole chiave e dei riassunti di ciascun testo che facilitano il processo di selezione di un testo da inserire nel proprio corpus, piuttosto che un altro.

Inoltre, la struttura della pagina del database include campi per filtrare i documenti simili a quelli presenti nella pagina iniziale, in particolare con "SOURCE", si permette di selezionare il sito da cui il documento è stato estratto, e con "SUPPRIMER" si procede per la rimozione individuale di un documento.

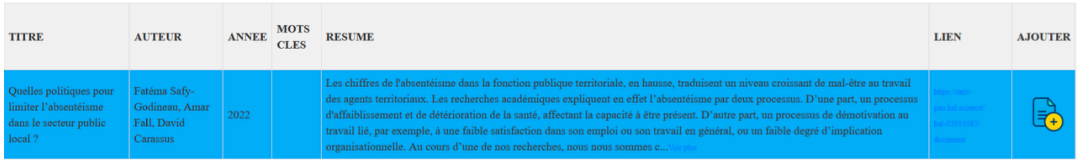


Fig. 2.12. Esempio di un documento selezionato per essere aggiunto al database, che subisce l'evidenziazione in blu nell'interfaccia *Accueil*.

Fig. 2.13. Esempio dei risultati e dei comandi nell'interfaccia *Base de données*.

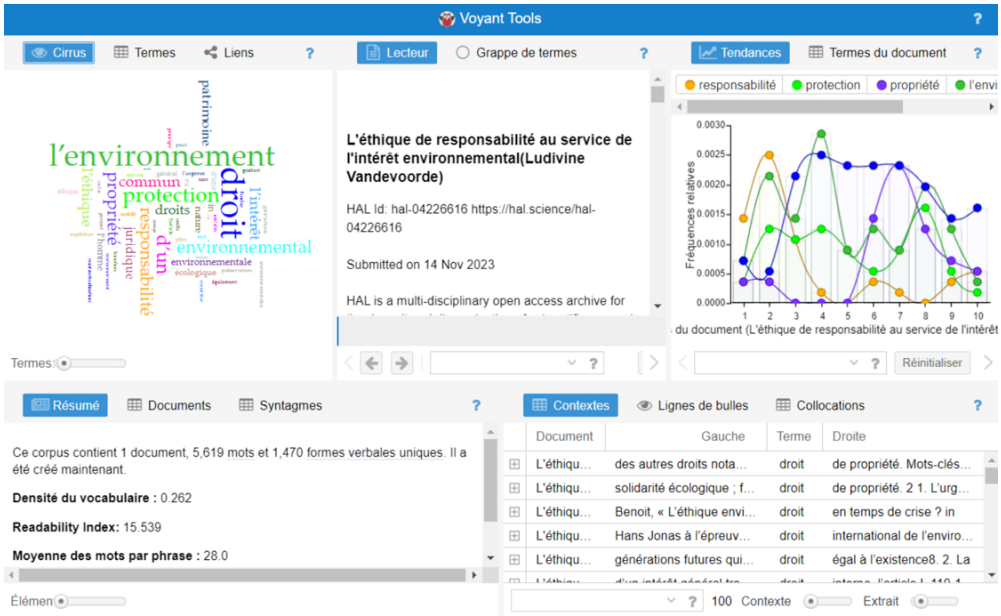
Nell'interfaccia Base de données/Database, è inoltre possibile scegliere come trattare il corpus virtuale, ovvero selezionare una delle quattro funzionalità principali che offre il software (Fig. 2.17.):

— l'esportazione dei documenti in formato .txt: questa funzionalità è particolarmente utile per inserire il corpus in software avanzati per l'estrazione automatica; infatti, alcuni siti web di analisi di corpora, come *TermoStat*, richiedono i testi in formato .txt, raccolti in un unico file. Pertanto, per evitare

che l'utente debba farlo con altri strumenti online, il che potrebbe aumentare i processi ed il carico di lavoro, il software offre la possibilità di esportarli direttamente in un unico file in formato .txt: dai documenti contenuti nel database, vengono recuperati tutti quelli con un collegamento ad un file PDF, vengono convertiti in un file di testo .txt e successivamente allegati tutti in un unico file. Una volta fatto questo, il file di testo creato viene scaricato sul computer dell'utente;

— l'esportazione dei documenti verso l'applicazione web Voyant-Tools[136];

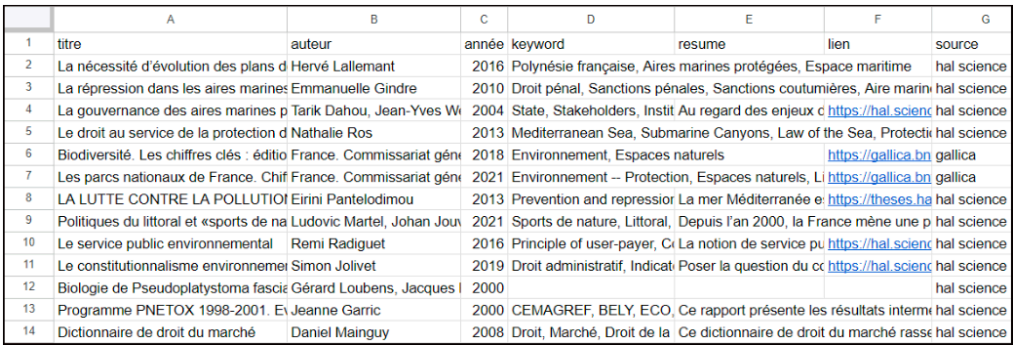
Fig. 2.14. Esempio di importazione diretta dei link ai PDF dei documenti selezionati del database sul sito *VoyantTools*.



[136] Sulla base della metodologia adottata dalla collega francese (Darricades 2023), abbiamo implementato questa funzionalità che in prospettiva futura ci può permettere di differenziare il modus operandi, sperimentando un'analisi basata principalmente sulle *mot-clés*. Il pulsante di esportazione, utilizzando la libreria Selenium, apre una pagina di Voyant-tools ed incolla dinamicamente tutti i link ai documenti PDF presenti nel nostro database (Fig. 2.14.).

Fig. 2.15. Esempio di risultato dell'analisi dei documenti selezionati dal database sul sito *VoyantTools*.

— l'esportazione dei metadati dei documenti in formato .csv, ovvero schedati in un file Excel: al fine di archiviare i dati di un corpus creato e renderli reperibili successivamente, permettendo di eliminarli dal software per effettuare una nuova ricerca, è disponibile un pulsante di esportazione CSV: i documenti CSV possono essere aperti con strumenti come Excel per visualizzarli in una tabella, in cui vengono inseriti tutti i dati contenuti nel database, associando ogni valore di ciascun elemento al relativo ID di colonna (Fig. 2.16);



	A	B	C	D	E	F	G
1	titre	auteur	année	keyword	resume	lien	source
2	La nécessité d'évolution des plans d	Hervé Lallemant	2016	Polynésie française, Aires marines protégées, Espace maritime			hal science
3	La répression dans les aires marines	Emmanuelle Gindre	2010	Droit pénal, Sanctions pénales, Sanctions coutumières, Aire marin			hal science
4	La gouvernance des aires marines p	Tarik Dahou, Jean-Yves W	2004	State, Stakeholders, Instit	Au regard des enjeux c	https://hal.scienc	hal science
5	Le droit au service de la protection d	Nathalie Ros	2013	Mediterranean Sea, Submarine Canyons, Law of the Sea, Protecti			hal science
6	Biodiversité. Les chiffres clés : éditio	France. Commissariat gén	2018	Environnement, Espaces naturels		https://gallica.bn	gallica
7	Les parcs nationaux de France. Chif	France. Commissariat gén	2021	Environnement -- Protection, Espaces naturels, Li		https://gallica.bn	gallica
8	LA LUTTE CONTRE LA POLLUTIOI	Eirini Pantelodimou	2013	Prevention and repressior	La mer Méditerranée e	https://theses.ha	hal science
9	Politiques du littoral et «sports de na	Ludovic Martel, Johan Jouv	2021	Sports de nature, Littoral,	Depuis l'an 2000, la France mène une p		hal science
10	Le service public environnemental	Remi Radiguet	2016	Principle of user-payer, C	La notion de service pu	https://hal.scienc	hal science
11	Le constitutionnalisme environnemer	Simon Jolivet	2019	Droit administratif, Indicat	Poser la question du cc	https://hal.scienc	hal science
12	Biologie de Pseudoplatystoma fascia	Gérard Loubens, Jacques I	2000				hal science
13	Programme PNETOX 1998-2001. E	Jeanne Garric	2000	CEMAGREF, BELY, ECO,	Ce rapport présente les résultats interm		hal science
14	Dictionnaire de droit du marché	Daniel Mainguy	2008	Droit, Marché, Droit de la	Ce dictionnaire de droit du marché rass		hal science

Fig. 2.16. Funzionalità "Esporta in formato .csv": estrazione automatica dei metadati in un file Excel.

— l'eliminazione totale dei documenti dal database, con una finestra di conferma per evitare cancellazioni accidentali.

Fig. 2.17. Principali funzionalità di estrazione delle conoscenze del software (dall'interfaccia in lingua italiana).

2.3.3 Risultati e prospettive future dell'implementazione informatica

L'implementazione di *CorpusBuilder* ha rappresentato un passo significativo verso l'ottimizzazione delle metodologie di raccolta e gestione dei dati, nell'ambito dell'analisi terminologica di un dominio di specialità. I risultati ottenuti da una prova di integrazione di questo strumento informatico nel nostro studio hanno confermato la validità di un approccio interdisciplinare, il cui contributo principale risiede nella capacità di supportare in modo efficiente le fasi preliminari di costituzione di un nuovo corpus o le fasi di arricchimento testuale di uno già

esistente: in particolare, sull'impronta del presente studio dottorale, una volta superate le limitazioni causate dalla raccolta manuale, il software *CorpusBuilder* permette di estendere la copertura del corpus a nuovi sottodomini, come il diritto ambientale marino e le specie marine invadenti, ampliando così la base testuale per l'analisi terminologica di futuri studi.

In primo luogo, la funzionalità di *web scraping,* che consente di acquisire in modo sistematico e massivo documenti da fonti online selezionate (quali HAL-Science, EUR-Lex, Gallica, *et al*.), conduce ad un notevole incremento del volume e della varietà dei testi, arricchendo il corpus con materiale pertinente ai nuovi sottodomini identificati. Parallelamente, la fase di classificazione, che include l'estrazione automatica di metadati quali titolo, autore, anno di pubblicazione, tipologia testuale e sottodominio, garantisce una strutturazione del corpus che facilita enormemente le conseguenti operazioni di filtraggio, selezione e analisi dei testi. Difatti, l'interfaccia utente di *CorpusBuilder* è stata concepita per rendere il corpus più accessibile e consultabile: la sistematizzazione della raccolta e della classificazione dei metadati permette alla ricerca terminologica da un lato di concentrarsi su aspetti più qualitativi e complessi del linguaggio specialistico, dall'altro di beneficiare di un ambiente dati più robusto e facilmente interrogabile. Ne consegue che la possibilità di visualizzare i testi unitamente ai loro metadati facilita la navigazione e l'individuazione di documenti specifici, contribuendo a una migliore comprensione del contesto d'uso dei termini.

In secondo luogo, le diverse modalità di estrazione dei dati dal corpus, sebbene calibrate sulle esigenze del presente studio, rappresentano approcci efficienti per avviare successive analisi terminologiche: a partire dall'estrazione in formato .txt, che consolida tutti i testi in un unico file, questa funzionalità riduce significativamente i tempi di compilazione e formattazione del corpus; l'estrazione e l'importazione del corpus in strumenti che consentono una rappresentazione visiva e concettuale, chiara ed immediata, del corpus, offrendo una panoramica ricca di elementi utili per una prima esplorazione dei dati; infine, l'estrazione strutturata dei metadati, ora automatizzata e standardizzata in un file Excel, costituisce la vera "chiave di volta" dell'implementazione del software, garantendo

una maggiore affidabilità e rapidità nella gestione delle informazioni essenziali per ogni documento.

In conclusione, l'implementazione del software da un lato ha permesso di sperimentare un'agevolazione delle fasi più tortuose di uno studio basato sull'analisi di un corpus testuale, attraverso l'integrazione di strumenti informatici avanzati nella ricerca linguistica, dall'altro, tramite l'interfaccia utente, ha condotto al miglioramento dell'accessibilità e della navigabilità di un corpus, come fonte di risorse divulgative.

Sebbene il percorso di sviluppo abbia incontrato le tipiche sfide di una progettazione di un software (come la scelta della metodologia di *scraping*, la gestione degli errori, le restrizioni di compatibilità con sistemi operativi Apple), il raggiungimento degli obiettivi iniziali apre ad importanti prospettive future. Il software, pur essendo stato concepito per le esigenze specifiche della presente tesi, possiede una flessibilità intrinseca che ne consente l'applicazione a un più ampio spettro di progetti di raccolta e analisi di dati testuali, data la varietà di documenti recuperabili dalle fonti utilizzate. In futuro, si potrebbe considerare l'ulteriore sviluppo di *CorpusBuilder* per includere funzionalità di analisi linguistica avanzate, direttamente integrate, o per espandere la sua compatibilità con ulteriori formati e tipologie di risorse. L'obiettivo ultimo rimane quello di affinare la capacità di *ZooCor* di essere una risorsa dinamica e completa per la terminologia del patrimonio naturalistico marino, supportando la ricerca e la divulgazione in modo sempre più efficace.