Università
Ca'Foscari
Venezia

Scuola Dottorale di Ateneo
Graduate School

Dottorato di ricerca
in Scienze del Linguaggio
Ciclo 28°
Anno di discussione 2017

Le variazioni terminologiche in un corpus giuridico parallelo italiano-arabo: studio linguistico-computazionale

SETTORE SCIENTIFICO DISCIPLINARE DI AFFERENZA:
L-LIN/01 GLOTTOLOGIA E LINGUISTICA

Tesi di Dottorato di Fathi Hassan Ahmed Fawi, matricola 955994

Coordinatore del Dottorato
Prof. Alessandra Giorgi

Supervisore del Dottorando
Prof. Marina Buzzoni

Co-supervisore del Dottorando
Prof. Rodolfo Delmonte

Ringraziamento

Desidero ringraziare tutti coloro che mi hanno aiutato a portare a termine, nel miglior modo possibile, la presente tesi.

In modo particolare ringrazio la professoressa Marina Buzzoni e il professor Rodolfo Delmonte, per il loro indispensabile sostegno e le loro insostituibili osservazioni elagite durante tutte le fasi del lavoro.

Un grazie di cuore va, inoltre, al professor Alessandro Lenci, direttore del Laboratorio di Linguistica Computazionale (CoLing Lab) di Pisa per i suoi illuminanti e preziosi insegnamenti fornitimi durante il corso di linguistica computazionale tenuto da lui al Dipartimento di Filologia, Letteratura e Linguistica all'Università di Pisa.

Un ringraziamento doveroso va, inoltre, a Felice dell'Orletta, ricercatore presso l'Istituto di Linguistica Computazionale A. Zampolli (ILC) del Consiglio Nazionale delle Ricerche (CNR) di Pisa. A lui va particolarmente la mia gratitudine per avermi fornito gli strumenti utili per taggare i testi italiani.

Infine, i miei sentiti ringraziamenti vanno alla mia famiglia, ai miei professori e ai miei amici, a cui dedico questo lavoro.

Indice

Lista di abbreviazioni

- Teoria Generale di Terminologia (TGT)
- Trattamento Automatico del Linguaggio (TAL)
- Linguistic Data Consortium (LDC)
- Amnesty International (AI)
- Organizzazione delle Nazioni Unite (ONU)
- Organizzazione Internazionale del Lavoro (OIL)
- Statistical machine translation (traduzione automatica statistica) (SMT)
- Log Likelihood ratio (LLR)
- Mutua Informazione (MI)
- Target Language (lingua di arrivo) (TL)
- Source Language (lingua di partenza) (SL)
- Multi-word terms (termini composti) (MWTs)
- Nome (N)
- Aggettivo (A)
- Preposizione (p)
- Verbo (V)
- Avverbio (AVV)

Elenco dei simboli di traslitterazione del sistema Buckwalter

Al fine di facilitare la leggibilità dei testi arabi è stato adottato il sistema di traslitterazione Buckwalter[1], i cui simboli sono indicati nella tabella seguente:

Lettera araba	Traslitterazione
ء	'
أ	>
إ	<
ا	A
آ	|
ب	b

1 Si è scelto in questa tesi Buckwalter perché si tratta di un sistema di traslitterazione dell'alfabeto arabo che viene adottato maggiormente negli strumenti del TAL arabo. Cfr. http://www.qamus.org/transliteration.htm

ت	t
ث	v
ج	j
ح	H
خ	x
د	d
ذ	*
ر	r
ز	z
س	s
ش	$
ص	S
ض	D
ط	T
ظ	Z
ع	E
غ	g
ف	f
ق	q
ك	k
ل	l
م	m
ن	n
و	h
ؤ	p
ي	w
ى	&
ئ	}
ه	y
ة	Y

1. Introduzione

La terminologia, intesa come la disciplina scientifica nata per compilare, standardizzare, e studiare i termini all'interno dei linguaggi specializzati, si è sviluppata notevolmente a partire dalla seconda metà del Novecento, in risposta all'esigenza esponenziale di organizzare e normalizzare le conoscenze scientifiche portate dall'evoluzione tecnologica, sociale e scientifica che ha caratterizzato tutto il ventesimo secolo.

L'onnipresenza dei termini nella vita quotidiana degli utenti di una lingua, specialisti o meno, che li utilizzano per motivi riguardanti sia le attività comunicative che la rappresentanza della conoscenza scientifica, ha fatto sì che lo studio della terminologia abbia polarizzato recentemente l'attenzione dei ricercatori operanti nelle diverse discipline scientifiche. Questo interesse crescente da parte degli studiosi verso la terminologia si è rinvigorito significativamente grazie allo sviluppo avuto nel campo di linguistica dei corpora, che ha consentito cioè di studiare ed analizzare le terminologie nel loro contesto, nonché allo stabilirsi degli approcci socio-cognitivi e comunicativi che invitano a considerare la terminologia in una prospettiva descrittiva, rifiutando quindi il pensiero tradizionale della terminologia basato concretamente su ottiche prescrittive.

The use of machine-readable corpora and large data banks offers terminological research significant advantages over traditional systems that were in use until quite recently. The huge amount of data terminologists have available allows them to obtain well-founded information about terms and to have much other information at hand. On the whole, this provides a more solid foundation to the decisions terminologists have to make throughout the process, and gives terminological activity increased flexibility and the possibility of responding better to user groups.[2]

Risulta indubbia in questo senso l'influenza esercitata dall'evoluzione nelle scienze informatiche sulla terminologia per quanto riguarda sia la metodologia e il tipo di trattamento che le applicazioni che si possono

2 Cabré, M. T., *Terminology: theory, methods and applications*, Amsterdam, Benjamins, 1998, p.164

effettuare tramite l'intelligenza artificiale[3]. Da lì è nata quindi la denominazione "terminologia computazionale" che si riferisce all'analisi automatica dei linguaggi specializzati, e che comprende la creazione dei corpora, l'estrazione di termini, la classificazione dei documenti, e l'analisi morfologica, sintattica e semantica dei termini, ecc..

Secondo Cabré[4] lo sviluppo della moderna terminologia è passato per quattro periodi:

a) le origini (1930–1960): fra i lavori terminologici di rilievo di quest'epoca viene in primo luogo il contributo di E. Wüster cui si è disegnata poi la teoria generale di terminologia (TGT) che riconosce la terminologia come disciplina scientifica autonoma con dei propri principi e obiettivi ben precisi;

b) la strutturazione del campo (1960–1975): questa fase è caratterizzata dallo sviluppo di computer mainframe e delle tecniche di documentazione. Così sono nate le prime banche dati e si è cercato di elaborare approcci finalizzati a standardizzare le terminologie all'interno delle lingue;

c) il boom (1975–1985): in questo periodo si è avuta una proliferazione dei progetti terminologici a livello nazionale indirizzati a modernizzare la propria lingua. Inoltre la diffusione del computer personale ha portato un grande cambiamento nelle condizioni dei dati terminologici;

d) l'espansione (1985–presente): quest'era ha vissuto l'evoluzione del trattamento automatico del linguaggio umano che ha fornito ai terminologisti strumenti efficaci e risorse adatte alle loro attività terminologiche. Per di più, è aumentata in questo tempo la cosiddetta *industria linguistica* che comprende i servizi di traduzione, interpretariato, localizzazione, doppiaggio, ecc. in cui la terminologia possiede ovviamente un ruolo centrale. Sono cresciuti, del resto, gli scambi di informazioni a livello internazionale.

3 Ivi, p.162
4 Ivi, p.5

Secondo Antia un termine può essere considerato come "un simbolo, linguistico o meno, che identifica concetti […] i cui confini sono chiaramente circoscritti"[5]. È una definizione non tanto distante da quella di Sager che, distinguendo tra le unità lessicali che possiedono riferimento generale e quelle con una proprietà precisa all'interno di un soggetto specifico, definisce i termini come quelle voci lessicali che "sono caratterizzate da un riferimento speciale dentro una disciplina"[6].

Risulta chiara, quindi, la relazione all'interno di un discorso specializzato tra un termine e il concetto, che è un'unità di pensiero, di conoscenza e di cognizione per la rappresentazione mentale delle realtà[7].

In realtà i termini racchiudono, secondo la teoria comunicativa di Cabré[8], tre dimensioni fondamentali e responsabili della produzione nonché dell'uso delle unità terminologiche: componente cognitiva relativa al mondo di conoscenza, componente linguistica come strumento per trasmettere e comunicare queste entità di conoscenza, e infine una componente socio-comunicativa riguardante l'uso effettivo dei termini all'interno delle società scientifiche. Per acquisire la veste terminologica e avere una specificità nell'uso linguistico prendendo distanza cioè dal resto delle parole di un testo, le unità terminologiche devono seguire, secondo Cabré, le seguenti condizioni:

1. da un punto di vista cognitivo, un termine deve basarsi su un contesto tematico occupando una precisa posizione in una struttura concettuale da cui viene determinato il suo significato. Inoltre, questo significato deve essere fissato esplicitamente e diventare una proprietà dell'unità terminologica;

2. in una prospettiva linguistica, i termini sono unità lessicali e quindi sono soggetti alle stesse regole linguistiche adottate nella lingua in cui si trovano,

5 Antia, B.E., *Terminology and Language Planning: An alternative framework of practice and discourse*. Amsterdam/Philadelphia, Benjamins, 2000, p.96
6 Sager, J., *A Practical Course in Terminology Processing,* Amsterdam, Benjamins, 1990, p.19
7 Cfr. Antia, B.E. *Terminology and Language Planning*, op. cit., p.82
8 Cabré, M. T., *Theories of terminology Their description, prescription and explanation*, op.cit., p183

partendo dai procedimenti di formazione lessicale fino alle caratteristiche morfosintattiche, ecc.;

3. in un'ottica pragmatica i termini occorrono in discorsi specializzati a cui si adattano in base alle loro caratteristiche tematiche e funzionali.

Tuttavia, una volta creati e organizzati nei dizionari o nelle banche dati, i termini potranno non soddisfare tutte le esigenze degli utenti di una lingua che ci si affidano per comunicare i loro discorsi specializzati o tradurre i loro testi in altre lingue. Questo fenomeno sembra dovuto alla distinzione effettuata da Cabré tra la funzione rappresentativa e quella comunicativa delle terminologie. Mentre nella prima ci si interessa all'organizzazione e alla standardizzazione dei termini, nella seconda direzione si considerano le prove empiriche dell'uso delle terminologie all'interno delle situazioni comunicative. Nell'ambito della funzione comunicativa dei termini può succedere che un termine non designi sempre, come sostiene Mayer[9], in modo univoco un unico concetto specialistico oppure che un concetto possa essere designato da più di una denominazione, dando luogo al fenomeno delle variazioni terminologiche.

La presente tesi si propone di analizzare le variazioni terminologiche in un corpus parallelo italiano-arabo. L'idea della tesi è nata dall'assunto che anche i linguaggi specifici o settoriali presentano delle variazioni a livello lessicale come è il caso della lingua comune, contrastando così con la teoria generale della terminologia basata sul principio di monosemia e monoreferenzialità secondo il quale i termini non devono subire delle variazioni per evitare qualsiasi ambiguità comunicativa. In effetti, lo stato dell'arte dell'estrazione delle variazioni da corpora linguistici specializzati dimostra che le evidenze empiriche fornite dall'uso effettivo delle terminologie all'interno dei contesti provano che i termini specializzati possono avere delle varianti sia sul piano concettuale che su quello denominativo e sotto diverse forme: morfologiche, sintattiche, morfosintattiche o semantiche. I lavori svolti in questa direzione si

9 Mayer, F. "Sinonimia ed equivalenza". In Magris, M., et al (a cura di) *Manuale di terminologia*, Milano, Hoepli, 2002, p.115

possono classificare in due categorie: lavori interlinguistici che utilizzano corpora bilingui o multilingui, paralleli o comparati, per estrarre e analizzare le diverse forme di variazione terminologica e di questo tipo ricordiamo Barbagianni[10] e Carreño Cruz[11]; e lavori intralinguistici basati su corpora monolingui come i tentativi di Freixa[12] per descrivere le variazioni in testi catalani nel campo ambientale, e Jacquemin[13] per i corpora in francese.

Partendo dall'importanza assoluta di identificare le diverse variazioni in un discorso specializzato sia per i lavori di traduzione che per le applicazioni del trattamento automatico del linguaggio (TAL), soprattutto per migliorare la performance dei sistemi di traduzione automatica, il presente lavoro cerca di individuare le variazioni terminologiche all'interno di un corpus giuridico parallelo italiano-arabo. L'obiettivo generale del lavoro è estrarre, descrivere e analizzare le relazioni di variazione denominativa dei termini giuridici in italiano e in arabo. Il lavoro si occupa solo dei termini composti. A tal fine è stato creato un corpus che contiene testi in italiano con la loro relativa traduzione in arabo, specializzato nel campo dei diritti umani nel mondo. L'opzione per questo genere testuale risponde effettivamente alla disponibilità, in formato elettronico, dei corpora specializzati paralleli italiano-arabi nonché alla qualità di traduzione di questi testi. La novità del nostro lavoro consiste nell'adottare un approccio computazionale nelle diverse fasi del lavoro.

Nel primo capitolo del lavoro cerchiamo di discutere alcune verità sulla variazione nei discorsi specializzati in generale e nel linguaggio giuridico in particolare, presentando, a tal riguardo, i diversi punti di vista degli approcci coinvolti nel campo terminologico. Nel capitolo successivo vengono trattati i diversi procedimenti di formazione dei termini in entrambe le lingue, una fase

10 Barbagianni C., "Verso un modello di variazione terminologica: un'analisi della terminologia della gestione dei rifiuti in testi normativi". In *Quaderni di Palazzo Serra* 25, 2014

11 Iveth Carreño Cruz, S., *Analyse de la variation terminologique en corpus parallèle anglais-espagnol et de son incidence sur l'extraction de termes bilingue*, tesi di master, Université de Montréal, 2004.

12 Freixa, J., *La variació terminològica. Analisi de la variació denominativa en textos de diferent grau d'especialització de l'àrea de Medi Ambient*, Tesi di dottorato, Universitat de Barcelona, 2002

13 Jacquemin, C., "Syntagmatic and paradigmatic representations of term variation" In *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics* (ACL'99)

di grande rilievo per conoscere le caratteristiche linguistiche dei termini giuridici in italiano e in arabo. Il terzo capitolo descrive il corpus utilizzato nella tesi. In una prima parte si parla della costituzione dei corpora linguistici, passando dalle tipologie diverse dei corpora fino ai criteri che determinano questi corpora. Poi si descrivono i delineamenti del corpus della tesi, comprese le motivazioni della scelta dei documenti nonché le diverse tappe del trattamento del corpus. Nel quarto capitolo si utilizzano dei pattern morfosintattici caratterizzanti dei termini in ambedue le lingue per estrarre dei termini candidati che poi, tramite le misure statistiche, vengono verificati al fine di selezionarne solo i termini rilevanti. Oggetto del quinto capitolo è unificare, tramite le tecniche di allineamento a livello di parola, le due liste dei termini monolingui per creare delle corrispondenze di traduzione. Il sesto capitolo sarà dedicato all'estrazione e all'analisi delle variazioni dei termini.

Capitolo I: Le variazioni terminologiche

1.1. La variazione nella terminologia

Fino ad alcuni decenni fa il fenomeno di variazione linguistica, sia denominativa che concettuale, si studiava solamente in testi appartenenti alla lingua generale, mentre nei linguaggi specializzati o tecnici tale fenomeno era quasi ignorato o inesistente. Questo era dovuto, secondo Dury&Lervad[14], al pensiero secondo il quale la funzione principale dei linguaggi specifici è informare o comunicare informazioni e conoscenze su un soggetto specializzato, il che richiede, quindi, una massima chiarezza. Questa necessità di non creare ambiguità comunicativa ha portato a credere che nei discorsi specializzati non ci debba stare spazio a fenomeni come sinonimia, polisemia, o altre forme della variazione terminologica. Tuttavia, contrariamente a questo pensiero tradizionale sulla terminologia, sono stati proposti recentemente tanti approcci descrittivi che condividono l'affermazione che i termini nei linguaggi settoriali sono soggetti alle stesse variazioni delle parole generali[15].

In un lavoro empirico, Daille et al.[16] presentano in dettaglio i diversi modelli di variazione terminologica e propongono un insieme di regole formative da applicare sia nella generazione che nell'identificazione di varianti nel testo. Gli studiosi vedono che i linguaggi specifici rappresentano un ambiente conveniente per le variazioni terminologiche, e sostengono la loro ipotesi con esempi tratti da un corpus medico dove un solo termine come *Epithelial cell* presenta più di 15 varianti a livello dello stesso corpus. E in un altro studio Daille fornisce una statistica secondo la quale le variazioni terminologiche in un testo sono stimabili tra il 15% e il 35%, in base al dominio testuale e ai tipi di variazioni analizzate[17].

---

14 Dury, P., Lervad, S., "La variation synonymique dans la terminologie de l'énergie: approches synchronique et diachronique, deux études de cas". In *LSP&Professional Communication* , vol. 8, n. 2, 2008

15 Cfr. Fernández Silva, S., Kerremans K., "Terminological variation in source texts and translations: A pilot study". In *Meta: Translators' Journal*, 56(2), 2011, p.319

16 Daille, B. et al., "Empirical observation of term variations and principles for their description". In *Terminology* 3(2), 1996

17 Cfr. Daille B., "Variations and application-oriented terminology engineering". In *Terminology* 11:1, 2005, p.163

Giunti a questo punto sarebbe opportuno soffermarci a chiederci cosa si intenda per variazione terminologica. In Daille et el. viene fornita la seguente definizione della variazione denominativa:
"A variant of a term is an utterance which is semantically and conceptually related to an original term"[18].

Dalla suddetta definizione risultano tre verità importanti riportate dagli autori:

1. il termine "utterance" indica che una variante di un termine deve essere una forma attestata tramite i corpora linguistici;

2. "original term" rivela che una variante terminologica viene determinata in riferimento ad un termine originale definito come "an authorised term either listed in a thesaurus or in a terminological resource";

3."semantically and conceptually related": in base a questa assunzione si possono dare varie interpretazioni, nel senso che la relazione semantica in questo caso può riguardare o un'equivalenza semantica con il termine originale (sinonimia) o un solo aspetto di distanza semantica dal termine di riferimento, oppure si riferisce ad un altro termine connesso concettualmente al termine originale[19].

La definizione della variazione in terminologia fornita da Freixa non è tanto differente da quella precedente. Secondo Freixa nel campo terminologico per variazione denominativa[20] si intende la presenza di diverse denominazioni per indicare un unico concetto con la condizione che tali denominazioni siano

18 Daille, B. et al., "Empirical observation of term variations", op cit, p. 201
19 Cfr. Daille B. "Variations and application-oriented terminology engineering", op cit, p.164
20 Fernández-Silva et al. distinguono tra due tipi di variazione denominativa: variazione denominativa senza conseguenze cognitive e variazione denominativa accompagnata da qualche effetto cognitivo. Si ha il primo caso quando si utilizzano forme diverse per designare un unico concetto senza variazione semantica come per es. *marine product* vs *sea product*; mentre l'altro caso si realizza quando l'uso di due (o più) forme differenti per indicare un unico concetto conduce ad alcuni mutamenti a livello semantico, differenze cioè che sottolineano caratteristiche proprie del concetto in questione, come per es. *marine product* vs *fishing product.* Fernández-Silva, S., et al., "Multiple motivations in the denomination of concepts: the case of "production area" in the terminology of aquaculture in French and Galician". In *Terminology Science and Research*, 2009, Vol.20. Visto il 13 Marzo, 2016, http://iitf.fi/cms/component/remository/func

lessicalizzate (escluse cioè quelle denominazioni basate sulle definizioni o sulle parafrasi) e ci sia un "minimum of stability and consensus among the users of units in a specialised domain"[21].

Nell'ambito del TAL il compito di riconoscere e estrarre le variazioni dei termini ha acquisito recentemente una particolare importanza, attirando l'attenzione dei ricercatori. Accanto alla sua rilevanza negli studi socio-linguistici che cercano di analizzare, a livello sia intralinguistico che interlinguistico, l'evoluzione diacronica o sincronica dell'uso linguistico, il riconoscimento delle variazioni denominative all'interno dei corpora linguistici ha un impatto diretto sulla performance di certe applicazioni del TAL, come l'*information retrieval*, la creazione delle ontologie, la traduzione automatica, ecc..

1.2. Approcci della terminologia

La Teoria Generale della Terminologia (TGT), elaborata dagli studiosi di Wüster[22] in base alle idee di quest'ultimo, è fra i primi tentativi scientifici indirizzati a riconoscere la terminologia come una disciplina scientifica autonoma. La TGT è stata idealizzata principalmente con lo scopo di descrivere e organizzare le conoscenze terminologiche per poter poi standardizzare i linguaggi scientifici, rendendoli strumenti di comunicazione più efficaci senza nessun rischio di ambiguità comunicativa[23].

Le caratteristiche della TGT si possono riassumere nei seguenti punti:

- la sua rigida considerazione della semantica dei linguaggi specializzati, vista

21 Freixa, J., "Causes of denominative variation in terminology: A typology proposal". In *Terminology*. 12(1), 2006, p.51

22 Wüster, E., *Introduction to the General Theory of Terminology and Terminological Lexicography*, Vienna, Springer, 1979

23 Cfr. Cabré, M. T., "Theories of terminology, their description, prescription and explanation". In *Terminology* 9(2), 2003

tramite un'ottica che la distacca dalla sua veste linguistica, ovvero i termini che esprimono i concetti semantici:

Within this framework concepts were viewed as being separate from their linguistic designation (terms). Concepts were conceived as abstract cognitive entities that refer to objects in the real world, and terms were merely their linguistic labels.[24]

- il principio di monoreferenzialità e univocità dei termini: la precisione e la specificità degli aspetti semantici dei linguaggi scientifici, un obiettivo ispirato, tra l'altro, all'esperienza di Wüster come ingegnere, hanno portato a credere che i termini debbano essere monosemici e univoci, il che significa che ogni termine indica solamente un concetto preciso;

- la visione limitata dei termini: la TGT considera i termini scientifici solo dal punto di vista lessicale, senza prestare attenzione ad altri aspetti linguistici o extra-linguistici delle unità terminologiche;

- Il lato sincronico di ogni approccio all'analisi terminologica[25].

Come si può notare la visione meramente prescrittiva della TGT la rende distante dalla situazione reale ed effettiva della terminologia, il che porta Cabré a scrivere che "Wüster developed a theory about what terminology should be in order to ensure unambiguous plurilingual communication, and not about what terminology actually is in its great variety and plurality."[26]

Le principali critiche mosse alla teoria classica della terminologia provengono, secondo Cabré, da tre discipline scientifiche:

1- le scienze cognitive: nell'ambito della dimensione cognitiva la conoscenza generale non va separata da quella specializzata in quanto la prima contribuisce all'acquisizione della seconda. Analizzando la dimensione

24 Faber Benítez , P., "The Cognitive Shift in Terminology and Specialized Translation". In *Monografías de Traducción e Interpretación,* Universitat de València, 2009, p.111
25 Cfr. Cabré, M. T, "Theories of terminology, their description, prescription and explanation", op cit., p.166.
26 Ivi, 167

cognitiva della terminologia, Sager sostiene che la struttura della conoscenza non è un "entità assoluta" bensì rispecchia lo stato di conoscenza contemporaneo e innovativo degli individui o dei gruppi di specialisti:

In their effort of determining the terms relevant to a subject, terminologists start from the analysis of limited domains of knowledge and build up complex systems of concepts which eventually intersect and overlap. A characteristic feature of this work is the difficulty of fixing the structure of knowledge at any one time because conceptual systems are relatively fluid entities constantly undergoing change, especially in the research and development of innovative science and technology. Consequently the terminologist has to be a subject specialist himself or have very close contacts with subject specialists in order to keep track of innovation in concepts and terminology respectively[27].

2- le scienze del linguaggio dove si contesta la rigida divisione tra il linguaggio generale e il linguaggio specializzato dal momento che entrambi i tipi di discorso si possono integrare. Dal punto di vista linguistico, oggi i termini non sono considerati, come afferma Sager, come singoli oggetti in dizionari o parte di linguaggi "semi-artificiali" sprovvisti deliberatamente da qualsiasi funzione degli altri elementi lessicali. In effetti, ciò è diventato possibile grazie allo sviluppo costante della linguistica dei corpora che ha consentito lo studio dei termini nell'ambito di ambienti contestuali e situazioni comunicative il che ha fatto acquisire ai termini nuove assunzioni teoriche nonché ulteriori metodi di compilazione e di rappresentazione[28].

3- le scienze di comunicazione che propongono modelli comunicativi in cui i discorsi specializzati nonché la loro rappresentazione sociale sono una parte integrante di ogni processo comunicativo al cui successo contribuiscono effettivamente i termini.

Così la terminologia, soprattutto dopo lo sviluppo avvenuto nella sociolinguistica e in particolare nei linguaggi specializzati nonché nel trattamento automatico del linguaggio che ha offerto degli strumenti efficaci e

27 Sager, J.C., *A Practical Course in Terminology Processing*. Amsterdam: John Benjamins, 1990, p.13
28 Ivi, p.58

risorse varie per studiare e descrivere i termini tramite i corpora, ha dovuto subire una specie di revisione sul versante sia teorico che pratico[29]. Questa revisione si è ispirata fondamentalmente alle nuove scuole di pensiero, e in particolare la Socio-terminologia, la Teoria Comunicativa della Terminologia, e la Terminologia Socio-cognitiva, che cercano di rappresentare alternativi paradigmi della concezione terminologica. Anche se invitano a studiare la terminologia da ottiche diverse, corrispondenti praticamente alle discipline scientifiche da cui provengono, queste scuole hanno in comune tante caratteristiche tra cui Peruzzo ricorda, oltre al rifiuto del metodo prescrittivo wusteriano, l'adozione dell'approccio descrittivo che prova a estrarre evidenze empiriche fornite appositamente da corpora testuali, nonché l'utilizzo dell'approccio semasiologico che può essere integrato con alcuni contributi onomasiologici[30].

1.2.1 Approccio socio-terminologico

L'approccio socio-terminologico invita a considerare le dimensioni sociali e pragmatiche per lo studio dei fenomeni terminologici. Su una base descrittiva, questo nuovo approccio considera essenzialmente il funzionamento dei termini nell'ambito del loro uso concreto e contestuale, cioè nella loro dimensione discorsiva e interattiva con altre discipline scientifiche anziché delimitare la conoscenza terminologica in un livello teorico.

Apparue sous la double influence de la sociolinguistique théorique et de la sociolinguistique de terrain, la socioterminologie se fixe comme object l'étude de la circulation des termes en synchronie et en diachronie, ce qui inclut l'analyse et la modélisation des significations et des conceptualisations. Elle possède une dimension sociocritique comme toute sémantique du discourse dans la mesure où elle relie la production de sens des termes avec les conditions de leur apparition. La

29 Cfr. Pelletier, J., *La variation terminologique: un modèle à trois composantes*, Philosophiæ doctor (Ph.D.), Université Laval, 2012, p.7

30 Peruzzo, K., *Terminological Equivalence and Variation in the EU Multi-level Jurisdiction: A Case Study on Victims of Crime*. Doctoral thesis in Interpreting and Translation Studies, IUSLIT, University of Trieste, 2013, p.116

circulation des termes est envisigée sous l'angle de la diversité de leur usages sociaux, ce qui englobe à la fois l'étude des conditions de circulation et d'appropriation des temes, envisagée comme des signes linguistiques, et non comme des étiquettes de concepts.[31].

All'interno di questo approccio "variationniste" delle terminologie si studiano: la descrizione degli usi effettivi dei termini; il rapporto fra i termini e il contesto linguistico, pragmatico e sociale; lo studio dei fenomeni della variazione linguistica, l'interdisciplinarità e la circolazione dei termini e dei concetti; e il rispetto delle diversità culturali e linguistiche [32].

Secondo Boulanger i termini, facendo parte di un sistema sociale sottoposto ad un continuo e insistente cambiamento, non possono cedere il loro diritto a variare sia semanticamente che lessicalmente.

Plutôt que de reconnaître la polysémie naturelle et la pertinence de la synonymie, on cherchait à retirer au terme son droit à la variation, à la fois em ce qui regarde le aspect sémantique (la polysémie) et en ce que a trait à la variation lexicale (la synonymie). Bien entendu, ce réductionnisme lexical était recherché; il était que l'effort d'"univocisation" avait pour objectif de ramener la multiplicité des situations et des variations de communication à une situation singularisée et simplifiée au possible[33].

1.2.2. Teoria comunicava della terminologia

Contro la teoria classica della terminologia, Cabré propone una teoria comunicativa della terminologia (TCT), adottando un trattamento multidimensionale (cognitivo, linguistico, pragmatico).

Terminological units have to be studied in the framework of specialised

31 Gaudin F.. "La socioterminologie". In *Langages*, 39e année, n°157, 2005, p.90
32 Pelletier, J., La variation terminologique: un modèle à trois composantes, op cit, p.10
33 Boulanger, J.C., "Présentation: images et parcours de la socioterminologie". In *Méta*, XL, 2, 1995, p.196

communication, which is characterised by such external conditions as sender, recipient and medium of communication, by conditions of information treatment, such as a precise categorisation determined externally by the conceptual structure, fixation and validated by the expert community, by specific and contextualised treatment of the topic, and, finally, by conditions which restrict the function and objectives of this communication" [34].

Secondo Cabré[35], la terminologia è un "campo interdisciplinare" che integra contributi della a) *teoria della conoscenza* che riguarda il modo di concettualizzare la realtà nonché i tipi di questa concettualizzazione e il rapporto dei concetti con le loro possibili denominazioni; b) *teoria della comunicazione* che descrive le tipologie di situazioni comunicative che possono verificarsi e la spiegazione delle caratteristiche e delle possibilità dei diversi sistemi di espressione dei concetti; e c) *teoria del linguaggio,* cioè la rappresentazione delle unità terminologiche nel linguaggio naturale.

In base alle finalità di creare risorse terminologiche Cabré distingue tra due funzioni principali della terminologia: funzione rappresentativa e funzione comunicativa. Nel primo caso le unità terminologiche si utilizzano effettivamente per denominare o rappresentare certi concetti corrispondenti a una situazione reale. Anche se comprende, direttamente o indirettamente, qualche attività comunicativa, la terminologia in questo caso ha un'unica finalità, cioè "to reach at a point where every «thing» has a «name», and even more, that every «thing» has «only one name»[36]". Nell'altro caso, invece, la finalità di una terminologia è comunicare la conoscenza contenuta sia tramite un canale comunicativo diretto, cioè tra gli specialisti, che in una maniera indiretta attraverso i mediatori comunicativi. Mentre nella sua funzione rappresentativa la terminologia appare fondamentalmente simbolica, utilizzando accanto alle unità linguistiche altri segni non appartenenti alla lingua naturale, tipo nomenclature, simboli, figure, ecc., e permettendo una

34 Cabré, M.T., "Theories of terminology. Their description, prescription and explanation. op cit, p.188
35 Cabré, M. T., *La terminología: Representación y comunicación. Elementos para una teoría de base comunicativa y otros artículos*, Barcelona, Institut Universitari de Lingüística Aplicada, 1999 , p.122
36 Cabré, M.T., "Standardization and Interference in Terminology". In *The Changing Scene in World Languages. Issues and challenges*, Amsterdam/Philadelphia, John Benjamins Publishing Company, 1997, p.54

relazione forma-concetto del tipo one-to-one, con la prospettiva comunicativa la terminologia appartiene principalmente alla lingua naturale aderendo di conseguenza alle sue regole e condizioni, tra cui è la possibilità che un unico concetto venga espresso in vari modi.

Secondo questa teoria, la terminologia è una sorta di processo di comunicazione, con diversi gradi di esperienza e di specializzazione, che varia non solo in base all'oggetto di comunicazione bensì ad altri fattori derivati dagli usi e dalle circostanze comunicative.

Terminology, conceived as a lexical variety mainly based on the subject of communication, presents an inner variation according to the geographical and the generational characteristics of speakers (at this point, we should add the choice of a particular conception of the subject) and depending on the degree of formality and the abstraction level in which communication is developed[37].

1.2.3 Teoria socio-cognitiva della terminologia

La teoria socio-cognitiva della terminologia, proposta da Temmerman[38], si basa sulla potenza cognitiva della terminologia nei linguaggi specifici nonché sulla variazione terminologica dovuta ai diversi fattori contestuali[39]. In un confronto tra la teoria tradizionale della terminologia e la teoria socio-cognitiva, Temmerman mette in rilievo i punti più salienti della sua nuova teoria della terminologia soprattutto nei riguardi degli elementi del *triangolo semantico*: *mondo, linguaggio* e *mente umana.*

- mentre il *mondo* nella teoria classica è visto come un oggetto, nel pensiero socio-cognitivo il mondo della scienza e della tecnologia è sperimentale ed è presente parzialmente nella mente umana:

37 Ivi, 57
38 Temmerman, R., "Questioning the univocity ideal. The difference between sociocognitive Terminology and traditional Terminology". In H*ermes. Journal of Linguistics* 18, 1997
39 Faber Benítez, P., "The Cognitive Shift in Terminology and Specialized Translation", op cit, p.16

Much of what we know and understand about the world is embodied, is the result of our sensory perceptions. It should be added that the other part is the result of our reasoning, which is interactive with the input via sensory perception, and via the transfer of other language users' ideas which we take in via discourse (written and spoken) for which language is the medium.[40].

- nell'ambito della teoria socio-cognitiva della terminologia il linguaggio non è considerato solamente per la sua capacità di denominazione, come è il caso nella teoria classica, bensì ha una funzione cognitiva (ideativa) accanto ad un'altra funzione testuale e comunicativa (interpersonale), il che vuol dire che il linguaggio ha un ruolo essenziale nel capire il mondo e nella comunicazione delle categorie.

- infine contro il mediocre ruolo attribuito nella TGT alla mente umana, che appare limitato alla facoltà di classificare gli oggetti in base al riconoscimento delle caratteristiche in comune, nell'interpretazione socio-cognitiva l'uomo non solo ha la capacità di concepire il mondo oggettivo ma riesce anche a creare in mente delle categorie prototipiche. Nel suo modo di concepire il mondo, la mente non può trascurare, per di più, la dimensione del linguaggio.

Contro la concezione tradizionale secondo la quale un termine, per questioni relative all'ambiguità comunicativa, non deve presentare né variazione lessicale (sinonimia) né variazione semantica (polisemia), la teoria socio-cognitiva della terminologia vede, invece, che la variazione terminologica si configura come un tratto funzionale nei linguaggi specializzati perché appare come conseguenza dell'evoluzione del significato. Ne consegue che non si può negare l'aspetto diacronico riscontrabile chiaramente nella continua discussione sulla denominazione delle cose nel mondo e sui loro significati.

Polysemy is functional in LSP [linguaggi specializzati] discourse, it is a consequence of meaning evolution. The constant discussion over how to name and what words mean is in the discourse of a community and has a time aspect.

40 Temmerman, R., "Questioning the univocity ideal", op cit., p.54

Polysemy is the result. What is univocal at one time may grow into polysemy depending on the type of category and how it is understood.[...]Synonymy is functional in LSP discourse; it reflects different perspectives[41].

1.2.4. Approccio testuale della terminologia

Partendo dall'analisi del discorso scientifico e tecnico, l'approccio testuale si interessa soprattutto all'analisi e allo studio delle terminologie attraverso gli ambienti contestuali, soprattutto dopo lo sviluppo evidente realizzato nella linguistica dei corpora. Secondo Condamines studiare terminologia tramite il discorso, cioè il contesto linguistico e situazionale, aiuta primariamente a individuare le lacune tra il discorso e le informazioni da trasmettere e, conseguentemente, a suggerire metodi per prevedere tali difficoltà[42].

La terminologie textuelle, dont le refus du référentialisme est plus ou moins marqué selon les écoles, déplace la problématique de la terminologie aux relations entre signifiés et à la spécificité du fonctionnement des signifiés dans les textes à caractère technique et scientifique; elle s'appuie essentiellement sur les méthodes de la linguistique de corpus pour proposer des listes de candidats termes, sans a priori ontologique[43].

Secondo Slodzian[44] il metodo testuale, che l'autore ritiene molto adatto alle esigenze e ai problemi relativi alla produzione "sfrenata" dei documenti specializzati, sconvolge le priorità: di fronte alla visione paradigmatica della teoria classica, l'approccio testuale ha le seguenti caratteristiche:
- si interessa al funzionamento reale delle unità lessicali in contesto;
- è un approccio descrittivo dei testi nonché delle unità lessicali (termini) che li compongono, e quindi è lontano dall'approccio normativo caratteristico della teoria classica:

41 Ivi, p.67
42 Condamines, A., "Variations in terminology: Application to the management of risks related to language use in the workplace". In *Terminology*, Vol. 16(1), 2010, p.44
43 Slodzian, M., "La terminologie, historique et orientations". In IC - 17emes Journees francophones d'Ingenierie des Connaissances, Jun 2006, Nantes, France. p.2
44 Slodzian M., "L'émergence d'une terminologie textuelle et le retour du sens". In *Le sens en terminologie*, Lyon, Presses Universitaires de Lyon, 2000, pp. 61-85.

- parte dai documenti testuali per creare ontologie terminologiche;
- analizza i termini attraverso il confronto tra i corpora diversi.

Questo approccio risulta, come sostiene Neveau, adatto allo studio delle variazioni terminologiche e a misurare la distanza tra la forma lessicale e il suo uso effettivo in contesto, in quanto "procede induttivamente a partire da occorrenze" in corpora specialistici[45]. Inoltre, grazie allo sviluppo nel trattamento automatico del linguaggio che ha consentito la disponibilità in formato elettronico di enormi quantità di documenti testuali monolingui e multilingui, l'approccio testuale riesce a mettere in evidenza le variazioni terminologiche sul piano sia intralinguistico che interlinguistico.

45 Neveu F., "Un aspect de l'apport des corpus à la terminologie linguistique: l'alignement". In D. Blampain, et al., *Mots, Termes, et Contextes,* actes des journées scientifiques du réseau Lexicologie, Terminologie, Tradution de l'Agence Universitaire de la Francophonie, Paris, Editions des Archives Contemporaines, 2006, p.382

1.3. Cause delle variazioni terminologiche

Molti ricercatori hanno studiato le cause delle variazioni nelle terminologie da diversi punti di vista. Cerchiamo in quanto segue di riassumere il contributo di Freixa[46] che ha cercato di passare in rassegna le varie cause che possono portare a questo fenomeno.

Come preliminari cause della variazione nei termini Freixa mette la ridondanza linguistica e l'arbitrarietà del segno linguistico.
La ridondanza linguistica, definita come "la possibilità di denominare qualcosa (idea, concetto, ecc.) tramite forme linguistiche differenti"[47], contribuisce certamente ad arricchire il sistema lessicale di una lingua consentendo cioè a multipli segni linguistici riferenti ad unico concetto di coesistere nonché di essere utilizzati egualmente dagli utenti della lingua.

Partendo dal principio saussuriano riguardante la natura arbitraria del segno linguistico, Freixa sostiene che la natura convenzionale della relazione tra il significante e il significato nel triangolo semiotico appare responsabile di casi come la sinonimia e la polisemia nonché del processo di concettualizzazione del reale, anche se ammettere l'esistenza della convenzionalità in questo senso non può negare la presenza di casi "motivati" o sistematici di segni linguistici come è il caso per es. di onomatopee o fono-simbolismo.

A language is a system which is intrinsically defenceless against the factors which constantly tend to shift relationships between signal and signification. This is one of the consequences of the arbitrary nature of the linguistic sign.[48]

In base ad una classificazione delle aree che motivano le variazioni denominative Freixa riconosce le seguenti categorie di cause:

46 Freixa, J. "Causes of denominative variation in terminology", op cit.
47 Ivi, p.54
48 Saussure, F., *Course in General Linguistics*. Translated by Roy Harris, Bloomsbury, London, 1983, p.87

1.3.1. Cause dialettali

La variazione terminologica creata a causa della provenienza diversa degli specialisti è stata sempre presa in considerazione da parte dei terminologisti.

Freixa suddivide le cause dialettali in tre sottoclassi:

- variazione geografica

La variazione per motivi geografici si manifesta chiaramente in quelle lingue parlate in più di un paese (come l'inglese, lo spagnolo, il francese, ecc.), in quanto per usi regionali si possono creare varie denominazioni per lo stesso referente, non solo sul piano della lingua generale ma anche nei linguaggi specializzati. Tuttavia, non tutte le lingue speciali subiscono lo stesso livello di variazione geografica, in quanto quelle più vicine alla comunicazione quotidiana sono più soggette a tale tipo di variazione.

- variazione cronologica

Questa tipologia di variazione si può realizzare attraverso il progresso della conoscenza nella vita umana, cosicché con il passar del tempo si crea una coesistenza tra un termine antico e un altro moderno, ognuno dei quali si riferisce allo stesso referente. Si tratta qui di un tipo di variazione che contrasta con la teoria wustriana della terminologia basata, come è stato detto precedentemente, sulla concezione sincronica della terminologia.
Come ricorda Condamines, l'evoluzione della conoscenza responsabile di quel tipo di variazione dipende a sua volta dall'evoluzione del contesto:
"The evolution of terms is heavily dependent on the evolution of the context: it is crucial to anticipate changing contextual elements (new needs, new methods, interdisciplinarity...)"[49].

49 Condamines, A., Variations in terminology, op cit, p.35

- variazione sociale

In quanto gli specialisti appartengono a diverse sfere professionali, le differenze sociali tra di loro possono avere qualche riflesso denominativo nei loro discorsi specializzati.

1.3.2. Cause funzionali

Le cause funzionali della variazione sono connesse agli usi, agli utenti della lingua e alle condizioni comunicative determinate da alcuni parametri come il canale comunicativo, il *topic* e il fine della comunicazione, e il livello della specializzazione del ricevente. In generale i discorsi specializzati sono meno soggetti ai fattori funzionali rispetto ai discorsi generali.

1.3.3. Cause discorsive

Si tratta in generale di variazioni retoriche e stilistiche motivate maggiormente dal desiderio di evitare le ripetizioni, di creare coesione testuale, o di realizzare un'economia linguistica. Inoltre, le cause discorsive si possono avere per "fattori soggettivi", riguardante cioè la ricerca di una certa originalità o espressività da parte dello specialista:

The author's tendency to colourfulness and diversity, to expressiveness and originality may result in his using unusual, striking colourful or contrastive expressions which are practically never true synonyms, even if they may occur in the field of terminology[50].

1.3.4. Cause interlinguistiche

Il contatto linguistico e culturale tra le altre lingue può essere una causa della

50 Irgl, V. "Synonymy in the language of business and economics." In Laurén, C. and M. Nordman (a cura di) *Special Language. From Human Thinking to Thinking Machines*, 1989, Philadelphia: Multilingual Matters LTD, p.278

variazione denominativa e soprattutto del fenomeno di sinonimia. Questo può avvenire tramite la tendenza ad adottare prestiti da altre lingue, anche se nella lingua d'arrivo non mancano i corrispondenti semantici. Secondo Freixa le motivazioni per quest'opzione potrebbero essere o per ragioni di prestigio oppure per fini di efficacia comunicativa, dal momento che "the most stable term is the one that originated in the language in which the concept was created or the one that has been accepted in international communication"[51].

1.3.5. Cause cognitive

Le cause cognitive derivano principalmente dalla diversità nel modo di concepire le entità nella realtà. Gli aspetti essenziali delle cause cognitive possono riassumersi in quanto segue:

- imprecisione concettuale, cioè la mancanza di esattezza riguardo ai contorni del concetto in questione, in quanto "the lack of conceptual stability is usually associated with denominative instability"[52];
- distanza ideologica che si potrebbe creare a causa della presenza di scuole di pensiero differenti che operano nello stesso campo di conoscenza, o per effetto delle scoperte scientifiche simultanee che portano talvolta gli ideatori a imporre le proprie denominazioni per fine di differenziazione ideologica;
-differenze della concettualizzazione: nell'ambito del processo di conoscenza si possono avere vari modi per strutturare e categorizzare la realtà. Questo può portare a rappresentazioni mentali dissimili e quindi a differenti concettualizzazioni.

One main variation in terminology is caused by the fact that several communities may use the same term but not exactly with the same meaning. This variation is due to what can be called "differences of point of view" and is related to the problem of multidimensionality, which in turn belongs to the realm of classification of concepts[53].

51 Freixa, J. "Causes of denominative variation in terminology: A typology proposal", op cit, p.63
52 Ivi, p.64
53 Condamines, A., "Variations in terminology: Application to the management of risks related to language use in the workplace", op cit, p.31. Va ricordato qui che il termine "multidimensionality" è stato menzionato

Secondo Montiel-Ponsoda et al[54] questo tipo di cause si incontra spesso in contesti multilingui dove diversi gruppi culturali e sociali potrebbero concepire e organizzare aree di conoscenza in maniere diverse in base a certe necessità o prospettive. In questo caso Whorf, discutendo la relazione tra il linguaggio e il pensiero umano, parla di relativismo nel modo di concepire il mondo:

"We are thus introduced to a new principle of relativity, which holds that all observers are not led by the same physical evidence to the same picture of the universe, unless their linguistic backgrounds are similar, or can in some way be calibrated"[55].

1.4. Classificazione delle variazioni terminologiche

Ci sono tante forme delle variazioni terminologiche. In un modello a tre componenti, Pelletier[56] suddivide le tipologie della variazione nei termini in tre categorie:

- variazione denominativa (VD), cioè l'esistenza di due (o più) denominazioni differenti che corrispondono allo stesso significato e allo stesso referente;

- variazione concettuale (VC), cioè un concetto può assumere diversi valori

---

da Bowker (Bowker, L., "Variant terminology: frivolity or necessity?" In 8th *EURALEX International Congress*. Liège: University of Liège, 1998, p.489): "We use the term multidimensionality to describe the phenomenon of classification that occurs when more than one characteristic can be used to distinguish between things, and hence those things can be classified in more than one way. A dimension represents one particular way of classifying a group of things; a classification with more than one dimension is said to be multidimensional". Secondo questa definizione di Bowker, la multidimensionalità, causata da un insieme di fattori, come lingua, cultura, tempo, livello differente di percezione, contesto, pensiero scientifico, ecc., significa che non tutti possono classificare un campo scientifico nello stesso modo, nonché può succedere che ogni volta un valutatore guarda lo stesso oggetto, lo vede con una prospettiva diversa, il che potrebbe avere il suo impatto non solo sulla formazione dei termini, bensì sul loro uso nel mondo di comunicazione.

54 Montiel-Ponsoda E., et al, "Multilingual variation in the context of linked data". In: *Proceedings of the 10th International Conference on Terminology and Artificial Intelligence,* 2013, Villetaneuse, Francia.

55 Whorf, B. L. , *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*, Cambridge, MIT Press, 1956, p.v

56 Pelletier, J., La variation terminologique, op cit.

secondo la concezione o l'uso che ne fanno i locutori, e in tal caso il concetto corrisponde allo stesso referente;

- variazione polisemica (VP), cioè una denominazione ha significati differenti e corrisponde a referenti diversi.

In quanto segue esponiamo le diverse forme della variazione denominativa:

1- variazione morfologica: che comporta modificazione di uno o più morfemi del termine originale attraverso la flessione, la composizione o la derivazione;
2- variazione sintattica: che riguarda i termini composti, in quanto comporta cambiamento all'interno della struttura sintattica del termine. Il cambiamento creato potrebbe risultare da: a) inserzione di un elemento all'interno delle componenti del termine composto; b) sostituzione delle parole funzionali, come le preposizioni, presenti nel termine; c) cambiamento dell'ordine delle parti interne del termine; d) coordinazione di una componente del termine ad altre unità lessicali; o e) riduzione, tramite l'omissione, della struttura compositiva dell'unità terminologica;
3- variazione morfosintattica: che concerne il cambiamento delle categorie grammaticali, o il trasferimento di un termine originale in parafrasi;
4- variazione semantica: che riguarda le relazioni semantiche e concettuali tra il termine di base e le varianti. La forma più frequente della variazione semantica è la sinonimia, cioè l'utilizzo di forme lessicali diverse per riferirsi allo stesso referente;
5- variazione ortografica che ha a che fare con la forma grafica dei termini.

Partendo dalla differenza della concettualizzazione proposta da Cabré, Aguado&Montiel[57]riconoscono tre gruppi fondamentali delle varianti terminologiche in base alla loro relazione con il concetto base cui si riferiscono:

1- varianti che coincidono semanticamente con lo stesso concetto, ma hanno

57 Aguado de Cea, G., Montiel-Ponsoda, E., "Term variants in ontologies". In *Proceedings of the 30th International Conference of AESLA*, Spain, Lleida, 2012

forme diverse, e in questo caso si parla di sinonimia assoluta[58]. Questo gruppo può comprendere le seguenti forme di varianti:
a- varianti grafiche e ortografiche (*localization* e *localisation*);
b- varianti di flessione (*cat* e *cats*);
c- varianti morfosintattiche (*nitrogen fixation* e *fixation of nitrogen*);

2- varianti che rispetto al termine base risultano, dal punto di vista sia semantico che formale, dissimili ma continuano, ciononostante, a riferirsi al principale concetto ontologico, sottolineandone probabilmente qualche aspetto riguardante lo stile, il registro oppure la specializzazione. In questo caso si parla di *sinonimia parziale*[59] oppure "terminological units that highlight different aspects of the same concept"[60].

Di questo gruppo fanno parte le seguenti forme di varianti:
a- varianti stilistiche o connotative (*man vs. bloke*);
b- varianti diacroniche (*tuberculosis vs. phthisis*);
c- varianti dialettali (*gasoline vs. petrol*)
d- varianti pragmatiche o di registro (*headache vs. cephalalgia*);
e- varianti esplicative (*immigration law vs. law for regulating and controlling immigration*);

3- varianti differenti sia semanticamente che formalmente anche se fanno riferimento a due concetti ontologici connessi. Le relazioni concettuali in questo caso possono essere di iponimia, di iperonimia o di meronimia.
Su un piano interlinguistico gli autori in un altro lavoro[61] suddividono le varianti di questo ultimo gruppo in due sottostanti tipologie: varianti verticali

58 Il concetto di sinonimia assoluta non è ben condiviso da tanti autori, dal momento che ogni variante possiede aspetti cognitivo-concettuali ben caratterizzanti. Pelletier per es. preferisce a "sinonimia assoulta" il termine "variazione denominativa" : "Nous sommes plutôt d'avis qu'il n'existe pas de synonymie absolue, mais bien des variantes dénominatives qui rendent compte des nuances propres aux situations de communication, qui, elles, sont multiples." Pelletier, J., La variation terminologique, op cit, p.27
59 Cfr. Cabré, M. T., "El principio de poliedricidad: la articulación de lo discursivo, lo cognitivo y lo lingüístico en Terminología". In *IBÉRICA* 16, 2008
60 Aguado de Cea, G., Montiel-Ponsoda, E., "Term variants in ontologies". op cit, p.25
61 Montiel-Ponsoda E. et al., "Multilingual variation in the context of linked data". op cit

e varianti orizzontali. Per le varianti verticali si intendono quelle che si riferiscono a concetti che presentano maggiori caratteristiche in comune anche se non sullo stesso piano di classificazione, come per esempio quando uno è più specifico dell'altro, come il termine *river* in inglese e *rivière* in francese. Le varianti orizzontali si hanno, invece, quando due termini comunicano concetti che condividono maggiori proprietà ma non in una maniera eguale, il che significa che un concetto può avere caratteristiche che l'altro non possiede e viceversa. Esempio di questo ultimo caso sono i due termini *Prime Minister* in inglese e *Presidente del Gobierno* in spagnolo.

Come sostiene Diki-Kidiri[62] appare fondamentale in queste classificazioni la distinzione tra le due nozioni *concetto* e *significato* dal momento che l'intersezione cognitivo-semantica tra i termini non va cercata nell'ambito del significato dei termini stessi ma sotto l'ombrella del concetto generale:

"La distinction du signifié et du concept permet de mieux situer les multiples perceptions particulières d'un même objet, perceptions culturellement motivées, et ce qui constitue la représentation de son unité ontologique indépendamment des visions particulières".

62 Diki-Kidiri, M., "Une aproche culturelle de la terminologie". In *Terminologie et diversité culturelle*, 2000

1.5. Equivalenza semantica nel dominio giuridico

Come tutti gli altri linguaggi specializzati, il linguaggio giuridico presenta delle peculiarità linguistiche adeguate alla natura della comunicazione giuridica. Una delle caratteristiche del discorso legale è la sua connessione assai forte con la lingua, non solo perché alla lingua spetta il compito di garantire la trasmissione del codice legale, bensì per la similitudine tra di loro: infatti tutti e due i sistemi sono basati su delle regole che garantiscono la costruzione della loro struttura e la loro evoluzione, nonché entrambi appaiono condizionati dalla dimensione sociale in cui sono inseriti, per cui la loro definizione dinamica degli oggetti va a pari passo con la continua evoluzione del contesto sociale[63].
Secondo Sabatini[64] i rapporti tra la lingua e il diritto presentano tre aspetti fondamentali:

- sia la lingua che il diritto sono creati dalla convenzione sociale come "istituti" primari per organizzare la vita sociale;
- entrambi i campi disciplinari presentano il carattere di "sistemi", caratterizzati cioè da una forte organizzazione interna;
- la "profonda consustanzialità tra la norma giuridica e la sua espressione linguistica" porta sempre ad un'opera di analisi del linguaggio.

Questa forte correlazione tra il discorso giuridico e il linguaggio ha determinato la necessaria collaborazione tra il giurista e il linguista. Mentre al primo spetta il ruolo di scrutare "i fatti di lingua negli aspetti che sono pertinenti alle teorie generali del diritto, all'interpretazione e all'applicazione delle norme", il secondo provvede a analizzare i testi giuridici, cercando di isolandoci "i tratti che li caratterizzano in quanto appartenenti a varietà di lingua distinte nel tempo, nella distribuzione geografica, nel mezzo di attuazione (scritto o parlato), nei registri[65]".

63 Tiscornia, D., Sagri, M, T., "Legal Concepts and Multilingual Contexts in Digital Information". In *Beijing Law Review*, Vol.3 No.3, 2012, p.74
64 Sabatini, F., "Analisi del linguaggio giuridico. Il testo normativo in una tipologia generale dei testi". In M. D'Antonio (a cura di) *Corso di studi superiori legislativi 1988-1989*, Padova, CEDAM, 1990, p. 675.
65 Mortara Garavelli, B., *Le parole e la giustizia: Divagazioni grammaticali e retoriche su testi giuridici*

Prima di trattare il concetto di equivalenza nel dominio giuridico vediamo opportuno soffermarci su alcune verità sulle lingue speciali.

Secondo Cortelazzo per lingua speciale (o linguaggio settoriale, linguaggio specialistico, sottocodice, lingua specifica, microlingua, ecc. per citare alcune delle varie denominazioni, riportate da Scarpa[66], date alle diverse varietà sociolinguistiche presenti all'interno di una lingua umana) si intende "una varietà funzionale di una lingua naturale, dipendente da un settore di conoscenze o da una sfera di attività specialistici, utilizzata, nella sua interezza, da un gruppo di parlanti più ristretto della totalità dei parlanti la lingua di cui quella speciale è una varietà, per soddisfare i bisogni comunicativi (in primo luogo quelli referenziali) di quel settore specialistico"[67].

Tuttavia, tali bisogni comunicativi non fanno di una lingua speciale un sistema linguistico chiuso o isolato, anzi l'interscambio, non solo lessicale, tra le lingue speciali e la lingua comune e perfino all'interno delle lingue speciali stesse è un fenomeno linguistico ben conosciuto. Il lessico delle lingue speciali proviene in misura maggiore dalla lingua comune sia in una maniera diretta, cioè senza modificazione semantica, sia tramite una rideterminazione semantica. Sul versante opposto non mancano nella lingua comune, soprattutto nell'epoca contemporanea, tanti vocaboli specifici derivati da lingue speciali, in un processo definito da Dardano "stilizzazione tecnologica"[68] e, nell'ambito del linguaggio giuridico, "un desiderio di nobilizzazione"[69]. Come causa principale di tale ingerenza massiccia dei termini specifici nella lingua comune (basti pensare alla statistica operata da Dardano[70] secondo la quale ben due terzi del lessico di una lingua provengono dalle lingue speciali) risulta la "forte carica espressiva dei termini specifici

italiani, Torino, Giulio Einaudi Editore, 2001, p.4

66 Scarpa, F., *La traduzione specializzata - Lingue speciali e mediazione linguistica*, Milano: Ulrico Hoepli, 2001, p.1

67 Cortelazzo, M.A., *Lingue speciali. La dimensione verticale*, Unipress, Padova, 1994, p.8

68 Dardano, M., "Profilo dell'italiano contemporaneo". In Serianni L., Trifone, P. (a cura di) *Storia della lingua italiana*, Torino, Einaudi, 1994, p.428.

69 Ivi, p.367

70 Ivi, p.310

una volta entrati nella lingua comune, sia quando perdono il proprio significato originale, acquistandone uno nuovo (...) che quando lo mantengono"[71].

Tradizionalmente le lingue speciali, per motivi di specificità concettuale, si caratterizzano dalla monoreferenzialità semantica, che non permette variazione funzionale che appare paradigmatica nella lingua di tutti i giorni. Tuttavia, gli approcci descrittivi della terminologia, come abbiamo visto nella parte precedente in questo capitolo, hanno dimostrato che l'univocità semantica nella terminologia è una visione idealizzata, ben lungi quindi dall'uso effettivo delle terminologie nel discorso comunicativo.

Il concetto dell'equivalenza semantica nel dominio giudiziale, soprattutto quello internazionale, è stato ultimamente tanto dibattuto sia da giuristi che da linguisti. I punti di vista delle due parti divergono riguardo all'esistenza o meno di un'equivalenza totale tra i sistemi giuridici diversi.
In un'ottica interlinguistica, Scarpa definisce l'equivalenza traduttiva come "la massima corrispondenza semantica, funzionale e socioculturale ottenibile tra testo di arrivo e testo di partenza tenendo conto della specifica situazione comunicativa in cui avviene l'attività traduttiva"[72].

Secondo Koller[73] l'equivalenza è un concetto relativo, determinato, cioè, non solo dalle condizioni storico-cultuali della produzione e della ricezione dei testi nella cultura di arrivo, bensì da un insieme di fattori linguistici e extra-linguistici, che sembrano talvolta anche contraddittori, quali concernono gli elementi essenziali dell'operazione di traduzione, e in particolare:
- il testo da tradurre: risultano di un certo rilievo qui le caratteristiche linguistiche, strutturali, stilistiche e estetiche;

71 Scarpa, F., La traduzione specializzata, op cit, p.17
72 Secondo Koller la traduzione è "il risultato di un processo di elaborazione testuale tramite cui un testo di una lingua di origine (source-language text) viene trasposto in un'altra lingua di destinazione (target-language text). Tra i due testi, quello di partenza e quello d'arrivo, si stabilisce un rapporto che può essere designato come di traduzione o di equivalenza", Koller, W., "The concept of equivalence and the object of translation studies". In *Target* 7, n.2., John Benjamin, Amsterdam, 1995, p.196
73 Ibidem

- la lingua di partenza e la lingua di arrivo: le norme linguistiche, stilistiche e estetiche;
- il traduttore: la propensione creativa, il suo modo di comprendere il testo, la teoria di traduzione adottata sia esplicitamente che implicitamente, i principi imposti dall'autore a proposito della traduzione, le linee guida del cliente nonché lo scopo dichiarato della traduzione, e, infine, le condizioni pratiche in cui si trova a lavorare;
- il lettore: le condizioni preliminari per la comprensione da parte del lettore della lingua di arrivo;
- il "mondo": la concezione del "mondo" tramite le varie classificazioni delle lingue, e le diverse realtà rappresentate dalle rispettive lingue.

Nell'ambito della teoria descrittiva della traduzione (Descriptive Translation Studies) la realizzazione di equivalenza traduttiva si può misurare secondo Koller in base ad un insieme di *modelli relazionli,* quali a) le circostanze extra-linguistiche trasportate dal testo; b) le connotazioni trasmesse dal testo tramite le modalità di verbalizzazione; c) le norme testuali e linguistiche che si applicano ai testi paralleli nella lingua d'arrivo; d) il modo di considerare il ricevente; e, infine, e) le caratteristiche estetiche del testo originale.

Target-language equivalents answer to *translational units* in the source text; both the similarities and the differences between the units of the source-language and their target-language equivalents result from the degree of which the values assigned to the relational frameworks are preserved[74].

Di relativismo tra i diversi sistemi concettuali parla anche Lakoff sostenendo che è difficile sottomettere tutti i sistemi di pensiero ad una determinata commensurabilità, realizzabile secondo Scarpa solo in caso di lingue culturalmente vicine[75]. In questo senso i criteri di commensurabilità riportati da Lakoff[76] sono:
- traduzione: due sistemi concettuali sono commensurabili se ciascuna delle

74 Koller, W. 1995, op cit, p.197
75 Cfr Scarpa, F., La traduzione specializzata, op cit, pp: 73-74
76 Lakoff, G. *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*, Chicago-Londoni, Univerisity of Chicago Press, 1987, p.322

lingue in questione può essere tradotta nell'altra conservando le condizioni di verità;

- comprensione: nel caso quindi che i due sistemi vengano compresi dalla stessa persona tramite la struttura preconcettuale della sua esperienza nonché la generale capacità di concettualizzare;

- uso: due sistemi risultano commensurabili se entrambi usano gli stessi concetti nello stesso modo;

- inquadramento: cioè quando ambedue i sistemi inquadrano situazioni allo stesso modo e, inoltre, la relazione di corrispondenza tra i loro concetti è del tipo one-one;

- organizzazione: ovvero quando i concetti all'interno dei due sistemi sono organizzati egualmente uno relativo all'altro.

Secondo Arntz nell'ambito della traduzione l'equivalenza terminologica si può realizzare qualora due termini condividano le stesse caratteristiche per creare una specie di "identità concettuale"[77].
Arntz riconosce i seguenti casi di equivalenza terminologica:

1-equivalenza concettuale completa: in questo caso tutte le sfere semantiche dei due termini sono interamente compresenti l'una nell'altra, come i seguenti esempi riportati dallo stesso autore: *private Daten* (tedesco) *informations personnelles* (francese);
2- sovrapposizione concettuale: in tal caso si realizzano due possibilità: a) l'intersezione di equivalenza tra i due termini è così ampia per tenere correlati i termini, come per es. *civil servant* (inglese) *Beamter* (tedesco); b) l'intersezione di equivalenza è tanto limitata per garantire una relazione di correlazione tra i due termini, come per es. *informatics* (inglese) *informatique* (francese);

77 Arntz R., "Terminological Equivalence and Translation". In Sonneveld H. B., Loening K. L. (a cura di) *Terminology. Applications in Interdisciplinary Communication*, Amsterdam, John Benjamins, 1993, p.13

3- inclusione: dove il concetto di un termine è compreso nel concetto dell'altro termine il quale potrebbe avere ulteriori caratteristiche rispetto al primo, come per es. *social* (francese) *sozial* (tedesco);

4- casi di similarità ortografica, o *faux amis*, che condurrebbero il traduttore a credere in un'equivalenza concettuale, mentre in realtà le sfere semantiche dei termini in questione sono ben differenti, come per es. *collège* (francese) *Kollegium*(tedesco);

In terminologia giuridica, si può parlare di equivalenza tra due termini in due lingue diverse quando questi due termini concordano in tutte le caratteristiche concettuali[78]. Tuttavia, i sistemi giuridici differiscono concettualmente da un ordinamento nazionale all'altro, e ne consegue una differenza nozionale tra i termini giuridici che compongono questi sistemi giudiziali. Questa realtà ha portato Sacco ad affermare che "I concetti creati, elaborati, definiti dal legislatore o dal giurista di un certo sistema non corrispondono necessariamente ai concetti elaborati per un altro sistema[79]". Esempio di questo parere è il concetto del *contract* in inglese che non trova corrispondente in francese perché il contratto francese non implica l'idea di una *consideration*.

Questo vuol dire che a livello interlinguistico in quanto la corrispondenza concettuale tra le diverse lingue riguardo alla realtà extra-linguistica è abbastanza limitata, la realizzazione del concetto dell'equivalenza assoluta appare quasi difficile se non impossibile. Non è strano quindi che ci siano pareri, come quello di Sandrini, che invitano alla revisione del concetto dell'equivalenza terminologica nel campo giuridico[80].

Secondo Sandrini l'equivalenza è definita sulla base delle corrispondenti caratteristiche concettuali che dipendono dall'intensione del concetto e la sua

78 Sacco, R., "Lingua e diritto". In *Ars Interpretandi* 5, 2000, p.125
79 Ivi, p. 126
80 Sandrini, P., "Comparative Analysis of Legal Terms: Equivalence Revisited". In *TKE '96,* Frankfurt: Indeks, 1996, p. 1

posizione nel sistema concettuale del campo scelto[81]. Dopo aver escluso l'equivalenza assoluta, Sandrini distingue tra due casi di equivalenza parziale:

1- due concetti possono avere in comune una sezione di corrispondenza, e in base alla dimensione di questa intersezione si può precisare la grandezza della potenziale corrispondenza, anche se rimane sempre la difficoltà di poter valutare in base all'intersezione tra due concetti, perché in alcuni casi una sola caratteristica potrebbe portare a due diversi concetti, mentre in altri casi un paio di caratteristiche differenti non possono condurre all'abbandono dell'equivalenza.

2- quando un concetto comprende un altro concetto, in tal caso si parla di relazione tra un concetto subordinato e un concetto sovraordinato.

Sandrini afferma infine che la terminografia in legge non può essere meramente la ricerca di identici concetti in due o più sistemi legali, perché questo approccio rischierebbe di coprire solo una minoranza dei casi. Quindi lui propone una metodologia che consideri solo l'equivalenza parziale o le caratteristiche di intersezione tra due termini, il che vuol dire l'abbandono dell'equivalenza assoluta in favore di un approccio comparativo più flessibile.

81 Un parere simile è condiviso anche da Sagri e Tiscornia secondo le quali "nei linguaggi specialistici, come quello giuridico, per stabilire una corrispondenza tra i termini di lingue diverse occorre individuare la corrispondenza tra concetti o istituti giuridici, in quanto la diversità di struttura dei diversi ordinamenti determina la difficile comparazione tra gli istituti presenti". Sagri, M. T., Tiscornia, D., "Le peculiarità del linguaggio giuridico. Problemi e prospettive nel contesto multilingue europeo". In *MediAzioni* 7, 2009, p.5. visto il 3/02/2016: http://mediazioni.sitlec.unibo.it

1.6. Sinonimia nella comunicazione specializzata

Come abbiamo visto nell'esposizione precedente i discorsi specializzati sono soggetti, come la lingua generale, ai diversi tipi di variazione denominativa. Fra le tipologie di variazione che hanno fatto discutere tanto nelle comunicazioni specializzate è la sinonimia. Mentre nell'uso generale della lingua la sinonimia è stata vista esclusivamente da un punto di vista stilistico, come elemento partecipante, cioè, nel creare la coesione lessicale di un testo[82], nei discorsi specializzati l'uso delle variazioni semantiche in generale e della sinonimia in particolare rappresenta spesso uno dei meccanismi comunicativi adottati.

In effetti, mentre l'equivalenza riguarda le relazioni cognitivo-semantiche tra i termini a livello interlinguistico, la sinonimia ha a che fare con l'entità semantica dei termini all'interno dello stesso sistema linguistico, anche se, come sostiene Mayer[83], tra tutti e due i concetti risulta, da un punto di vista metodologico, una specie di parallelismo: nella terminologia orientata alla traduzione per es. identificare le relazioni sinonimiche tra i termini a livello intralinguistico rappresenta una fase importante per evidenziare poi le equivalenze a livello lessicale esistenti tra le diverse lingue.

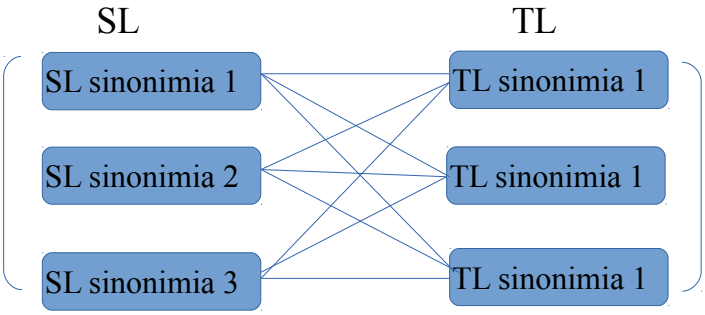


Fig.(1): Relazioni di sinonimia e corrispondenza in base al significato dei termini[84].

82 Cfr. Haliday, M. A. K, Hasan, R., *Cohesion in English*. London, Longman, 1976
83 Mayer, F. sinonimia ed equivalenza, op cit, p.116
84 Rogers, M., "Synonymy and equivalence in special-language texts. A Case Study in German and English Texts on Genetic Engineering". In Trosborg, A., (a cura di): *Text Typology and Translation*. Amsterdam: John Benjamins, 1997, p.218

Secondo Ullmann[85] due termini sono sinonimi se tra di loro si soddisfanno i criteri di intercambiabilità, cioè ogni termine può sostituire l'altro in tutti i contesti, e dell'equivalenza cognitivo-emotiva.

In base ai due suddetti criteri Lyons[86] riconosce quattro tipi di sinonimia:

a) sinonimia completa e totale: in questo caso si trovano soddisfatti entrambi i criteri, e qui si parla si sinonimia "assoluta" o "reale" che rappresenta effettivamente un caso raro nell'ambito delle relazioni semantiche tra le unità terminologiche;

b) sinonimia completa ma non totale, in cui è presente l'equivalenza cognitivo-emotiva, ma non l'intercambiabilità;

c) sinonimia totale ma non completa, dove esiste l'intercambiabilità ma manca l'equivalenza cognitivo-emotiva;

d) sinonimia non completa e non totale, dove mancano tutti e due i criteri.

Mentre la sinonimia completa e totale appare abbastanza rara se non impossibile, quella totale ma non completa, secondo la definizione precedente, risulta più probabile.

Più che equivalenza cognitivo-emotiva, Alhaj parla di contesto per indicare una relazione relativa di sinonimia:

"Synonyms is always related to context. Two lexical items are perfectly synonymous in a given context or in several contexts, but never in all contexts. The term used to describe this is relative synonymy.[87]"

85 Ullmann, S. *The Principles of Semantics*, Glasgow: Jackson and Oxford: Blackwell, 1957, p.108
86 Lyons, J. *Introduction to the theoretical linguistics*, London, Cambridge University Press, I968, p.448
87 Alhaj, A., *Understanding Semantics. A Textbook for Students of Linguistics and Translation*, Anchor Academic Publishing, 2015, p.20

Capitolo II: Formazione dei termini in italiano e in arabo

2.1 Il lessico giuridico

I termini, sia semplici che composti, sono soggetti in qualsiasi lingua umana agli stessi procedimenti formativi delle parole presenti nella lingua generale. In effetti, il linguaggio giuridico, come tutti gli altri linguaggi specializzati, deriva proprio dal linguaggio ordinario per arrivare poi a creare una sua peculiarità lessicale-semantica, o i cosiddetti i tecnicismi giuridici:

I linguaggi specialistici non sembrano presentare quelle limitazioni o semplificazioni rispetto alla lingua comune che talora sono state ipotizzate, ma sono dotate di tutte le potenzialità di natura lessicale, fonetica, morfosintattica e testuale tipiche della lingua comune. Tali potenzialità vengono regolarmente utilizzate (e, alcuni casi, addirittura iper-utilizzate) nella costruzione di testi specialistici[88].

Il linguaggio giuridico, adoperato nelle leggi, nei testi normativi e nelle scienze giuridiche, è in tutti in paesi di avanzata civiltà il frutto di una secolare opera di ricostruzioni parziali all'interno dei linguaggi naturali, ricostruzioni parziali incidenti principalmente sulla dimensione semantica dei linguaggi stessi: attraverso queste ricostruzioni il linguaggio giuridico è diventato un linguaggio tecnico, nel senso, soprattutto, di un vocabolario tecnico introdotto nella struttura di un linguaggio naturale[89].

Secondo Ralli[90] il lessico giuridico è caratterizzato da una natura molto eterogenea, costituito, accanto alle parole prettamente giuridiche o i tecnicismi specifici, di parole passate dalla lingua comune, dagli altri linguaggi settoriali, oppure di neologismi e prestiti da altre lingue, ecc.. In effetti, questa caratteristica di non avere confini precisi distingue il linguaggio giuridico dagli altri linguaggi settoriali, in quanto nel caso giuridico "vi rientra tutto ciò che può avere interesse per la vita associata degli uomini"[91]. Questa formazione "aperta" del linguaggio giuridico lo avvicina al lessico

88 Gotti, M., *I linguaggi settoriali*, Firenze, La Nuova Italia, 1991, p.7
89 Scarpelli, U., "Semantica giuridica". In *Novissimo digesto italiano*, vol.XVI, Torino, Utet, 1969, p.995
90 Ralli, N., "Terminografia e comparazione giuridica: metodo, applicazioni e problematiche chiave". In *InTRAlinea, online Translation Journal*, 2010, http://www.intralinea.org/specials/article/1727, cliccato il 30-4-2015
91 Serianni, L. *Italiani scritti, Bologna,* il Mulino, 2003, p.108

della lingua comune, anzi tra di loro, cioè il linguaggio giuridico e la lingua comune, succedono frequenti scambi di termini[92]. Quindi dal momento che il diritto regola la vita quotidiana delle persone, può succedere che non solo nel linguaggio del diritto si utilizzano termini provenienti dalla lingua comune, bensì "il parlante comune ha una certa facilità ad apprendere e riusare in senso metaforico e allusivo alcuni tecnicismi del diritto"[93]. Una volta ancorati nel nuovo ambiente giuridico, quei termini passati dal linguaggio ordinario acquisiscono, tuttavia, nuovi aspetti semantici, e ciò succede attraverso una rideterminazione semantica dei termini. Tuttavia, la natura aperta del linguaggio giuridico non permette facilmente di racchiuderlo entro una precisa definizione[94]

In linea generale la lingua del diritto presenta alcune peculiarità linguistiche che la distinguono sia dalla lingua comune che dalle altre lingue speciali. Fra le caratteristiche lessicali ricordiamo, oltre ai tecnicismi, la tendenza alla nominalizzazione, la presenza di termini arcaici e di forestierismi. Un'altra caratteristica lessicale del linguaggio giuridico che accomuna sia i tecnicismi che i termini passati dalla lingua comune è, in linea teorica, la tendenza alla precisione denotativa. Tuttavia, parlare di precisione non significa che tutti i termini giuridici devono essere monosemici, perché la variazione terminologica nei discorsi specializzati è diventata ormai una realtà, per cui Jori sostiene che "La polisemia esiste in larga misura nel linguaggio giuridico, quindi la monosemia rappresenta più che altro un obiettivo condiviso"[95].

2.1.1 I tecnicismi

Per i tecnicismi giuridici si intendono quei termini che, nel contesto giuridico, "non hanno nessun tasso di ambiguità, essendo parole che si usano solo nelle

92 Cfr. Belvedere, A., "Il linguaggio del codice civile: alcune osservazioni". In Scarpelli, U.&Di Lucia, P. (a cura di) *Il linguaggio del diritto* , Milano: LED, 1994, p.405.
93 Mantovani, D., "Lingua e diritto. Prospettive di ricerca fra sociolinguistica e pragmatica". In *Il linguaggio giuridico. Prospettive interdisciplinari*, Milano, Giuffrè, 2008, p.25
94 Cavagnoli, S., *La comunicazione specialistica*, Roma, Carocci, 2007, p.86
95 Jori, M., "Definizioni e livelli di discorso giuridico". In *Il linguaggio del diritto*, op cit., p. 371

rispettive accezioni tecniche: possono essere ignorate, ma non fraintese in un'accezione diversa"[96].
Questi sono alcuni esempi dei tecnicismi giuridici italiani riportati in Garavelli[97]:
- *anatocismo*: capitalizzazione degli interessi di una somma dovuta, mediante aggiunta al capitale degli interessi via via maturati;
- *anticresi*: contratto con cui il debitore o un terzo consegnano al creditore, a garanzia del credito, un immobile i cui frutti serviranno per il pagamento dell'interesse e del capitale;
- *laudemio*: tassa dovuta dall'enfiteuta al padrone;
- *sinallagma*: obbligazione reciproca che in un contratto vincola entrambe le parti a prestazioni corrispettive.

Di tecnicismi giuridici in arabo possiamo ricordare:
- jnHp: è un crimine punito dalla legge, con la reclusione per più di una settimana;
- Al<brA' : esenzione del debitore dal pagamento dei debiti;
- EryDp : richiesta rivolta per iscritto alle autorità pubbliche;
- AlmqAyDp: permuta
- wSAyp: tutela

Insieme ai termini esclusivamente e tecnicamente giuridici, il lessico giuridico si serve pure di termini attinti alla lingua comune, con modificazione del loro significato tramite un processo di ridefinizione, o *specializzazione semantica*[98], con l'assegnazione, cioè, di un significato specifico a parole d'uso comune; un significato, quindi, "che non coincide con quello col quale esse vengono normalmente adoperate"[99]. Esempio di questo tipo di rideterminazione semantica in italiano è il termine *colpa*: nel diritto penale il concetto di colpa "presuppone che il soggetto non abbia volontà di commettere il fatto, imputabile quindi a sua disattenzione od omissione"[100]. Nella lingua comune *colpa* viene usato, tuttavia, per indicare la

96 Serianni, L., *Italiani scritti,* op cit., p.81
97 Mortara Garavelli, B., *Le parole e la giustizia",* op cit, p. 10
98 Scarpa, F., La traduzione specializzata, op cit, p.44
99 Mortara Garavelli, B., *Le parole e la giustizia",* op cit, p. 11
100 Serianni, L., *Italiani scritti,* op cit., p.81

piena intenzione di fare qualche comportamento censurabile: “non è colpa mia! ecc.”.

Il processo di produrre nuovi significati da parole già esistenti nel vocabolario della lingua comune esiste anche nella lingua araba:

La langue commune écrite atteste una grande vitalité dans la mesure où elle a cédé beaucoup de mots aux langues de spécialité. Nous pouvons mème dire que tous les termes arabes sont des extensions des mots de la langue commune, l’arabe n'ayant ni grec ni latin”[101].

Esempio di questa specializzazione lessicale è il termine *AlznA* (adulterio) che nella lingua comune si riferisce alla relazione sessuale pre o extra-matrimoniale, mentre nel linguaggio giuridico riguarda solo la violazione della fedeltà coniugale.

Inoltre, ci sono i tecnicismi collaterali che sono dei termini “altrettanto caratteristici di un certo ambito settoriale, che però sono legati non a effettive necessità comunicative, bensì all’opportunità di adoperare un registro elevato, distinti dal linguaggio comune”[102], come per es. usare “attingere”, invece di “raggiungere”, ecc.. Tuttavia, non è sempre facile la distinzione tra i tecnicismi specifici e quelli collaterali. Secondo Serianni ci sono degli originari tecnicismi collaterali che, dopo un lungo uso tecnico nell’ambiente giuridico, hanno acquistato alcune caratteristiche dei tecnicismi specifici, diventando, cioè, insostituibili o almeno in rapporto rigido e stabile con la cosa designata. Ne è esempio il termine *delazione* “devoluzione, attribuzione a qualcuno dell’eredità di un defunto.”

Tutti i tecnicismi, sia specifici che collaterali, del linguaggio giuridico presentano delle soluzioni morfologiche eguali a quelle della lingua comune, nel senso che i procedimenti formativi di quei tecnicismi o terminologie specifiche sono soggetti alle regole di formazione delle parole operanti nella

101 Reguigui, A., *La Créativité Lexicale en Terminologie Arabe,* Ontario, Université Laurentienne, 1994, p.81
102 Ivi. p.82

lingua comune. Secondo Grossman&Rainer[103] "Si ritiene che una parola si riferisca a un concetto unitario, sia modificabile solo globalmente, e che eventuali parti costituenti siano inseparabili e presentino un ordine fisso". Ciò vuol dire che la parola, in base alla definizione precedente, rappresenta il nucleo di qualsiasi formazione terminologica. Tuttavia, tali procedimenti non sono solamente formali, bensì di natura semantico-formale attraverso il principio del cosiddetto *lessico mentale*[104], definito come l'insieme delle parole memorizzate dai parlanti nonché delle relazioni che i parlanti stabiliscono fra queste parole. È attraverso il principio del lessico mentale che si possono spiegare fenomeni strettamente connessi alla formazione delle parole come il *blocco* e *l'analogia*. Mentre il primo significa che una parola nuova non riscontra accettabilità presso i parlanti a causa dell'esistenza di un'altra parola sinonima ben radicata nella lingua (es. *rubatore:ladro*); il secondo concetto rimanda alla coniazione di neologismi sul modello di parole ben determinate nella memoria dei parlanti con particolare implicazione semantica (es. *giornalista squillo- ragazza squillo*). Nell'ultimo esempio possiamo osservare che il nome composto *giornalista squillo* non deriva solamente dalla regola di composizione N+N, ma specificamente dal modello *ragazza squillo*, di cui acquisisce l'implicazione semantica.

In quanto segue trattiamo in breve i diversi procedimenti della formazione dei termini in italiano e in arabo:

2.2 Formazione dei termini in italiano

2.2.1 La composizione

Secondo Dardano il procedimento formativo di composizione è "il grande serbatoio" da cui la lingua italiana moderna attinge per rinnovare e arricchire i suoi vocaboli.
Per la sua analiticità e per la sua rilevante- e in certo modo ordinata e programmabile- produttività, questo tipo di formazione [cioè la composizione] delle

103 Grossman, M, Rainer, F.(a cura di), *La formazione delle parole in italiano*, Tübingen, Niemyer, 2004, p 4
104 Ivi, p.7

parole si adatta alle esigenze di sempre nuove e articolate terminologie corrispondenti allo sviluppo e alla rapida penetrazione della tecnica del mondo di oggi[105].

I composti si formano da parole libere appartenenti sia a categorie grammaticali diverse che a unica categoria, rispettando, tuttavia, i canoni generali riguardanti la concatenazione delle categorie lessicali[106].

Per formare un composto, però, le due parole candidate devono
a) avere in comune una possibile relazione semantica;
b) produrre un complesso da un concetto unico;
c) non essere, separatamente, accessibili alle regole sintattiche, cioè non consentono nessuna modificazione all'interno di loro[107].

Le possibili combinazioni delle categorie lessicali in italiano generano un nome tranne che i due costituenti siano aggettivi[108], come dimostra il seguente schema:

$N + N \rightarrow N$ $[capo]_N + [stazione]_N \rightarrow [capostazione]_N$

$N + A \rightarrow N$ $[campo]_N\ [santo]_A \rightarrow [camposanto]_N$

$A + N \rightarrow N$ $[alto]_A\ [piano]_N \rightarrow [altopiano]_N$

$V + N \rightarrow N$ $[porta]_V\ [lettere]_N \rightarrow [portalettere]_N$

$V + V \rightarrow N$ $[sali]_v\ [scendi]_V \rightarrow [saliscendi]_N$

105 Dardano, M., *La formazione delle parole nell'italiano di oggi*, Roma, Bulzoni Editore, 1978, p.141
106 In questo senso si possono spiegare per esempio fenomeni come la non possibilità di composti costituiti da N+P, V+A, perché nel primo caso nella frase è la preposizione a precedere il nome, mentre nel secondo l'aggettivo è un modificatore del nome e non del verbo. Cfr. Grossman, M, Rainer, F.(a cura di), *La formazione delle parole in italiano*, op cit., p. 33
107 Ibidem
108 Secondo Scalise&Bisetto può produrre un aggettivo anche la concatenazione A + N, quando A è un aggettivo di colore e N è un nome che definisce il colore dell'aggettivo precedente, come nel caso rosso-mattone, Scalise, S. & Bisetto, A., *La struttura delle parole*, Bologna, Il Mulino, 2008, p.124

$P + N \rightarrow N$ $[dopo]_P [guerra]_N \rightarrow [dopoguerra]_N$

$A + A \rightarrow A$ $[grigio]_A [verde]_A \rightarrow [grigioverde]_A$

Non importa nei composti se tra le parti costituenti c'è una divisione grafica, cioè staccate oppure sono scritte in una forma attaccata. Il criterio principale, quindi, per decidere se l'unione di due (o più) parole rappresenta un composto o meno è il significato prodotto da tale unione rispetto al significato originario delle singole componenti: se la concatenazione produce un significato assolutamente nuovo rispetto a quello originario delle parti costituenti, in tal caso si parla di un composto, come per es. *tavola rotonda*, che non indica più un tipo di mobile rotondo ma una specie di convegno o riunione scientifica; se invece l'unione grafica di due elementi non crea un terzo significato nuovo e diverso, in questi casi non si tratta di composti, come per es. *dimmi*, dove le parti componenti conservano ancora i loro originari significati[109].

- I composti N + N

I composti di questo genere si possono classificare secondo i seguenti principali criteri:

- la relazione sintattico-semantica tra i costituenti del complesso. E qui si parla di distinzione tra c*omposti subordinati, composti coordinati e composti attributivi o appositivi*. Nei composti subordinati si istituisce tra le parole costituenti un rapporto semantico di subordinazione, come di specificazione: *capostazione* (capo della stazione), o attraverso una parafrasi: *pesce spada* (pesce con il muso allungato in forma di una spada appuntita)[110]. Nel rapporto di coordinazione, però, i due costituenti partecipano parallelamente a formare l'entità definita dal composto, come *caffellatte, chiaroscuro, bar pasticceria*

109 Nandor, B., "Sostantivi composti nell'italiano contemporaneo". In *Lingua Nostra*, 1978, XXXIX/4, p.117

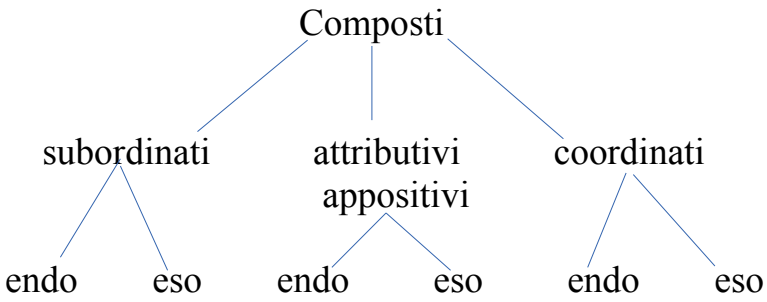
110 Grossman, M., Rainer, F., *La formazione delle parole in italiano* op cit., p.40

ecc.. Sono composti appositivi[111], invece, quelli in cui il nome testa è seguito da un altro nome con la funzione di apposizione, come *viaggio lampo, parola chiave, discussione fiume*, ecc. (dove si può notare qui la rideterminazione semantica del secondo elemento del complesso: *lampo*, cioè "di breve durata", *chiave* per significare "essenziale", *fiume* nel significato di "lungo").

- Endocentricità vs. esocentricità. Un altro criterio usato per classificare i composti N+N è la distinzione tra endocentricità e esocentricità, che rimanda alla nozione fondamentale di testa, definita come "il costituente iperonimico che, oltre alla categoria semantica, determina anche normalmente aspetti grammaticali della parola complessa come la categoria sintattica, il genere o la classe flessiva"[112]. I composti endocentrici[113] sono quelli in cui è possibile identificare un costituente con funzione di testa, come per es. *vagone letto, capostazione* (qui *vagone, capo* sono rispettivamente la testa dei complessi). Fanno, invece, composti esocentrici quelli in cui la peculiarità semantico-sintattica del composto non si attribuisce a nessuno dei costituenti, o in altre parole sono dei costrutti in cui manca una corrispondenza tra i tratti sintattico-semantici del composto e quelli dei costituenti come in *pellerossa* che indica una persona appartenente alla razza degli uomini dalla pelle rossa, o *portalettere* che non è un tipo di lettere ma una persona animata.

Concludendo la classificazione dei composti N+N possiamo citare qui lo schema classificatorio di Scalise&Bisetto dei composti[114]:

111 Secondo la classificazione di Grossman&Rainer (ibidem p.40) i composti appositivi, definiti *apposizionali*, fanno parte dei composti coordinati: "Sono interpretabili come coordinati perché l'oggetto cui fanno riferimento ha le caratteristiche, pur se in senso figurato, come detto, di entrambi i nomi costituenti".
112 Ivi, p.11
113 Secondo Scalise&Bisetto (op cit., p.124) si può identificare un costituente come testa di un composto quando tra di loro (cioè tale costituente e il composto) si stabilisce un rapporto di identità sia di categoria che di tratti sintattico- semantici. Per identificare la testa di un composto Scalise&Bisetto applicano il test "È UN", come per es. il composto *cassaforte*, i cui costituenti sono N+A = N. Per quanto riguarda la categoria lessicale il test si applica così: *cassaforte* "È UN" aggettivo o "È UN" nome? In quanto la risposta è "È UN" nome, quindi possiamo concludere che il costituente del composto con la categoria "nome" è quello che funge da testa del complesso (in questo caso *cassa*). Per quel che riguarda l'applicazione semantica del test, possiamo chiedere similmente *cassaforte* è "È UNA" *cassa* o "È UN" *forte*? Poiché *cassaforte* è una specie di cassa, quindi anche in questo caso si può concludere che l'elemento *cassa* è la testa semantica del composto.
114 Scalise&Bisetto, op cit. p.132



Rappresentano una sottoclasse dei composti N+N anche i composti nominali deverbali (definiti *compound-like phrases* da Scalise&Bisetto[115]). È un tipo di composti dove il nome-testa è un nome eventivo /di processo (process nominal)[116], derivato da un verbo transitivo o inaccusativo- intransitivo, mentre il nome modificatore può essere complemento oggetto diretto o indiretto del verbo (in)transitivo oppure soggetto del verbo inaccusativo[117].
- concessione permessi
- trasporto merci
- raccolta rifiuti

Tuttavia, questo tipo di composti presenta qualche peculiarità rispetto agli altri tipi di composti, cioè la possibilità di ammettere l'inserzione di altri materiali lessicali all'interno dei suoi costituenti, come per es. la modificazione aggettivale. Baroni et al riportano le diverse occorrenze di questo tipo di composti con la modificazione aggettivale:

- raccolta rifiuti tossici N[NA]
- raccolta diurna rifiuti [NA]N

115 Scalise, S., Bisetto, A., "Compounding: Morphology and/or syntax?" In Mereu, L.(a cura di) *Boundaries of morphology and syntax*, Amsterdam/Philadelphia:John Bejamins,1999, p.39
116 I nomi di processo (process nominal) occorrono sempre con i loro interni argomenti, e quando vengano seguiti da un solo argomento, questi viene interpretato necessariamente come il loro interno argomento. Cfr. Ouhalla, J., Shlonsky, U. *Themes in Arabic and Hebrew Syntax*, Kluwer Academic Pub; 2002, p.195
117 Cfr. Baroni, E. et al., "The dual nature od deverbal nominal constructions: Evidence from acceptability ratings and corpus analysis". In *Corpus Linguistics and Linguistic Theory* 5(1), 2009

- difficile raccolta rifiuti A[NN]
- raccolta nuovi rifiuti N[AN]
- raccolta diurna nuovi rifiuti [NA][AN]
- raccolta rifiuti efficiente [NN]A
- raccolta nuovi rifiuti tossici N[ANA]
- difficile raccolta rifiuti tossici A[N[NA]]
- raccolta efficiente nuovi rifiuti [NA][AN]
- raccolta nuovi pericolosi rifiuti N[AAN]
- difficile raccolta efficiente rifiuti tossici [ANA][NA]
- raccolta speciale efficiente rifiuti [NAA]N
- difficile raccolta rifiuti efficiente A[NN]A

Una gran parte delle occorrenze sopracitate viene attribuita, secondo gli studiosi, allo stile della scrittura dei titoli, soprattutto nel linguaggio dei giornali. La modificazione aggettivale risulta, inoltre, un punto di distinzione tra i composti nominali deverbali e quelli non deverbali:
- trasporto merci pericolose
- *treno merci pericolose[118]

- I composti V+N

In questo tipo di composti il primo posto viene occupato da un verbo mentre il secondo da un nome che funge generalmente da argomento interno diretto del verbo antecedente, quindi da complemento oggetto (a), oppure da soggetto in pochi casi (b):

a) macinacaffè	b) batticuore
attaccapanni	bollilatte
portaerei	marciapiedi

118 Delfitto, D., Paradisi, P.,"Prepositionless genitive and N+N compounding in (Old) French and Italian". In *Romance Languages and Linguistic Theory 2006 :Selected papers from 'Going Romance',* Amsterdam, 7–9 December 2006, John Benjamins, p.55

Secondo la classificazione endocentricità/ esocentricità il composto VN appare esocentrico, dal momento che nessuno degli elementi componenti del composto può fungere da testa del complesso: non lo è il primo elemento perché si tratta di un verbo[119] che è diverso dalla categoria grammaticale del complesso (cioè il nome), e non può esserlo anche la seconda parola perché, anche se si tratta di un nome, non assume le caratteristiche semantico-sintattiche del complesso, come per es. *portabandiera* che certamente non è un tipo di bandiera.

Dal punto di vista semantico, i composti di questo genere possono dare interpretazione agentivo/strumentale (*lanciasiluri, imbrattacarte, ecc*.); locativo (*marciapiedi, battiscopa, ecc.*); oppure eventivo o di "attività" (*alzabandiera, ammainabandiera, ecc*.)[120]. Anche se i composti VN danno tradizionalmente un output di un nome, non mancano nel frattempo alcuni usi referenziali o aggettivali del tipo V+N, come i composti *mangiafumo, lanciamissili, mozzafiato*, come per es. *un patto bloccaprezzi, ecc*..[121]

119 Ci sono dei pareri (Zuffi S., "The nominal composition in Italian. Topics in generative morphology". in *Journal of Italian Linguistics* 2,1981; Bisetto, A., "Note sui composti VN dell'italiano". In *Fonologia e morfologia dell'italiano e dei dialetti d'Italia : atti del 31. Congresso della Società di linguistica italiana*, Roma, Bulzoni, 1997) che sostengono che si tratta qui di un caso di nominalizzazione, cioè l'interpretazione nella maggior parte dei casi del verbo nella prima posizione del costrutto come un nome agentivo o strumentale (attraverso, cioè, il suffisso nominalizzante *-tore* che verrebbe poi cancellato come per es. *arricciacapelli*: arricciatore di/per capelli*, ecc.*), e che questo nome è la testa del complesso e quindi il composto appare endocetrico. "Da un punto di vista sincronico, l'idea che le formazioni VN siano in realtà costituite di due nomi sembra la più ovvia in ragione dell'interpretazione delle formazioni stesse. Produttivamente, infatti, sono possibili solo composti ad interpretazione agentivo/strumentale", Bisetto, A. *Note sui composti VN dell'italiano,* op cit, p.509. In conseguenza di tale ipotesi, cioè guardare alle formazioni VN come NN e quindi come formazione endocentrica, cambia anche la relazione tra i componenti del complesso: invece della relazione argomentale tra le parti del composto Bisetto (ivi:517) parla di relazione di modificazione: "Un nome che, nel lessico, si accompagna ad un derivato agentivo può solo essere interpretato come modificatore". Questo vuol dire che nei composti del tipo *accendigas, portasapone* i nomi *gas, sapone* non sono argomenti dei nomi *accenditore, portatore* che fungono da testa dei due composti.

120 Bisetto, A., "Note sui composti VN dell'italiano", op cit., p.509

121 Davide, R., "Al limite tra sintassi e morfologia: i composti aggettivali V-N nell'italiano contemporaneo". In Grossmann, M., Thornton, A.M. (a cura di) *La formazione delle parole. Atti del XXXVII congresso della Società di Linguistica Italiana*, Roma, Bulzoni, pp. 465-486, 2005

- I composti N+A

I composti nominali con un elemento aggettivale hanno la testa a seconda della posizione dell'aggettivo: se l'aggettivo è postnominale, la testa del complesso è a sinistra (*nave spaziale, camposanto*), e la testa a destra nel caso dell'aggettivo in posizione prenominale (*primo attore, francobollo)*. Il criterio definitorio per decidere se la sequenza NA rappresenta un composto o meno è la funzione attribuita agli aggettivi che "devono agire da restrittori del nome, non devono funzionare da aggettivi qualificativi"[122]. Questa funzione svolta dall'aggettivo è quindi di natura restrittiva, come in *cartone animato, terraferma*, *tavola rotonda*, ecc. ma non mancano, tuttavia, dei casi in cui succede una conversione per quanto riguarda i concetti di determinante e determinato tra il nome e l'aggettivo, come nei casi: *pettirosso, capinera*, dove l'aggettivo viene determinato dal nome che assume qui la funzione di complemento di limitazione.[123]

- I composti A+A

Questo tipo di composti crea una formazione di aggettivi coordinati, come *agrodolce, cristiano sociale, bianconero*. Invece di apparire nella loro usuale forma libera, gli elementi componenti del composto possono subire la cancellazione di vocale (*biancazzurro, socialdemocratico, imperial-regio*), o di una sequenza (*ferrotranviario,* dove è stato cancellato *viario*).

- I composti con elementi neoclassici

Non meno importanza possiedono nella lingua italiana e soprattutto nella terminologia tecnico-scientifica i composti con elementi neoclassici, di origine cioè greca o latina. "La composizione con elementi neoclassici è il tipo di formazione delle parole che utilizza elementi formativi tratti dalle lingue classiche per coniare termini di ambito tecnico-scientifico, usati

122 Bisetto, A., c*omposizione con elementi italiani.* In Grossman, M., Rainer, F. (a cura di), *La formazione delle parole in italiano* op cit., p.44
123 Nandor, B., *Sostantivi composti nell'italiano contemporaneo,* in: op cit, p.119

primariamente con funzioni designative e classificatorie"[124]. Dal greco ci sono per es. *antropo-, biblio-, crono-, emato-, -grado;* e dal latino*: -cida, igni, quadri-, -voro,* ecc..

Questo genere di composti presenta qualche peculiarità rispetto alle regole di formazione delle parole nella lingua italiana. I composti neoclassici sembrano diversi dai composti normali, dai derivati e dagli affissi. Non sono pienamente composti perché non sono formati dalla combinazione di due elementi liberi, come è il caso dei composti. Non sono neanche derivati in quanto non sono formati dalla combinazione di un elemento libero con un altro legato (affisso). Poi, diversamente dagli affissi, gli elementi formativi neoclassici non occupano sempre una posizione fissa, cioè iniziale o finale (*cronografo, isocrono*), e possono, inoltre, rappresentare basi di parole derivate, come per es. *antro- dendro-, antrosi, dendrite*[125], ecc..

2.2.2 La derivazione

La derivazione è il processo attraverso cui si formano parole nuove per mezzo dell'affissazione, cioè i prefissi e i suffissi, come per es. *contento → scontento, ubbidire → disubbidire, industria → industriale*, accettabile → *accettabilità,* ecc.. Con i prefissi, come si vede dagli esempi, i derivati mantengono la loro categoria grammaticale, mentre con i suffissi la categoria dei derivati cambia.

I suffissi, che rappresentano una maggioranza rispetto ai prefissi e che rappresentano un processo di formazione delle parole molto produttivo, possono formare nomi, aggettivi, verbi e avverbi a partire da nomi, verbi, ed aggettivi. Una classe importante dei suffissi sono quelli nominali che formano nomi a partire da verbi, nomi e aggettivi. Con i suffissi deverbali, cioè quelli che si applicano a basi verbali[126], si possono avere nomi di azione, cioè nomi

124 Antonietta Bisetto, "composizione con elementi italiani", op cit., p.69
125 Ivi, p.70
126 Scalise&Bisetto, La struttura delle parole, op cit, p.99

che indicano l'evento denotato dal verbo, e questo si fa di solito per mezzo di suffissi come *-zione (alienazione), -mento (abbinamento). -tura (cucitura), -aggio (decapaggio),* ecc.; e nomi di agente cioè nomi che indicano l'agente dell'azione del verbo con suffissi come *-tore (trasportatore), -(a/e)nte (cantante- assorbente),* ecc.. Poi ci sono i suffissi nominali denominali che formano nomi da basi nominali, come per es. *-aio (fioraio), -ista (flautista), -eria (libreria), - ificio (panificio),* ecc. Infine, ci sono anche i suffissi che formano nomi da aggettivi (*bellezza, idoneità*), aggettivi da nomi o da verbi (*storico, deperibile*), verbi da nomi o da aggettivi (*pietrificare, politicizzare*) o avverbi da aggettivi (*nascostamente*), ecc..

2.2.3 I latinismi e i forestierismi

Il lessico giuridico italiano è, come sostiene Jacqueline, "il risultato di una stratificazione di termini "entrati" in epoche diverse con la recezione dei diversi modelli di diritto straniero"[127]. Una caratteristica più evidente del linguaggio giuridico italiano è la presenza assai considerevole tra i suoi vocaboli di parole d'origine latina. I latinismi colti presenti nella lingua italiana in generale e nel linguaggio del diritto in particolare "hanno fatto il loro primo ingesso nel volgare medievale quando, da un lato, questo aveva già acquisito la propria fisionomia fonologica e grammaticale, e, dall'altro lato, il latino aveva cessato già da tempo di essere la principale lingua parlata e soprattutto non era più la lingua materna di nessuno"[128]. Rispetto agli altri linguaggi settoriali, il linguaggio giuridico riceve la maggior parte dei latinismi: secondo uno spoglio effettuato da Garavelli[129] sul Grande Dizionario Italiano dell'Uso di Mauro[130], dei 1.010 lemmi latini registrati nel dizionario, ci sono ben 163 lemmi che appartengono alla lingua del diritto, con una percentuale del 16,13%.

127 Visconti, J., "Prestiti e calchi: dove va la lingua giuridica italiana". In Bambi, F., Pozzo, B. (a cura di), *Dove va l'italiano giuridico*, Firenze, Accademia della Crusca, 2012, 185-194, p.186

128 Mantovani, D., Pellecchi, L, *La lingua del diritto:formazione, uso, comunicazione*, http://dsg.unipv.it/didattica/insegnamenti/la-lingua-del-diritto-formazione-uso-comunicazione.html, p.2 consultato il 1-6-2016

129 Mortara Garavelli, B., "Persistenza del latino nell'uso giuridico odierno". In *L'Accademia della Crusca per Giovanni Nencioni*, Firenze, Le Lettere, 2002, pp:423:433, p. 424

130 De Mauro, T., *Grande Dizionario Italiano dell'Uso*, Torino, UTET, 1999

Ci sono poi i prestiti dalle altre lingue, soprattutto dall'inglese e dal francese. Secondo Jacqueline[131]nella lingua giuridica italiana i prestiti inglesi provengono da almeno tre tipi di fonti: a) termini relativi a istituti proprio del *common law*, per motivi di prestigio o ispirazione da questo esercitati; b) fonti riguardanti l'internazionalizzazione della prassi contrattuale, soprattutto nella legislazione bancaria e finanziaria; c) la norma giuridica comunitaria e in questo caso si tratta di un inglese usato per veicolare i concetti di *civil law*.

Inoltre, una grande parte dei forestierismi francesi è stata trasmessa al linguaggio del diritto italiano durante l'epoca del dominio napoleonico in Italia e del trapianto dei Codici francesi[132].

2.3 Formazione dei termini in arabo

La lingua araba è una delle lingue semitiche più antiche del mondo. Oltre ad essere la lingua parlata in ventidue paesi arabi, l'arabo è anche la lingua del Corano, grazie al quale ha potuto sopravvivere fino ai nostri tempi. La domanda che si impone in questo caso è la seguente: come riesce una lingua come l'arabo, così antica nella storia, ad essere usata, a livello sia del parlato che dello scritto, fino ai nostri giorni, adattandosi a tutti i tipi di mutamenti sociali e di innovazioni scientifiche? La risposta a questa domanda è che l'arabo è caratterizzato da una struttura linguistica così flessibile da permettere ai propri meccanismi formativi di coniare nuovi termini per affrontare il grande sviluppo mondiale nella formulazione di nuovi concetti.

L'arabo non è una lingua rigida, in quanto fu creato per diventare lingua mondiale, con delle capacità formative adattabili a ogni situazione, senza che questo adattamento influisca sulle sue origini, le sue regole o le sue caratteristiche. Se gli esperti moderni non riescono a utilizzare l'arabo nelle loro attività moderne, questo è dovuto solo al fatto che loro (cioè gli esperti) hanno abbandonato la lingua araba e

131 Visconti, J., Prestiti e calchi, op cit, p.186
132 Cfr Mantovani, D.,"Lingua e diritto. Prospettive di ricerca fra sociolinguistica e pragmatica", op cit. p.39

i suoi paradisi recandosi ai misti deserti stranieri per esprimere le loro arti e le loro scienze trasportate dalle altre lingue[133].

L'aggiornamento linguistico dell'arabo nei confronti delle nuove realtà cognitivo-concettuali si caratterizza da due aspetti: interno e esterno. Il primo caso è successo nella tappa islamica della civiltà araba, quando l'arrivo dell'Islam ha comportato nuovi concetti religiosi che la lingua araba ha dovuto esprimere. L'altro lato, quello esterno, di quell'adattamento è avvenuto nell'era della rivoluzione industriale e tecnologica del tempo moderno quando l'arabo si è trovato costretto ad assorbire le moderne realtà concettuali avvenute nel mondo della scienza, soprattutto dopo lo stabilirsi della necessità di trasferire in arabo, attraverso le traduzioni dalle altre lingue, questi nuovi progressi scientifici e tecnologici.

In quanto segue esponiamo come si formano i termini nella lingua araba.

2.3.1 La derivazione

È un procedimento di formazione di nuove parole in arabo basato su radici linguistiche preesistenti nei dizionari della lingua araba (basi tematiche) e secondo delle regole derivazionali stabilite da parte dei linguisti e dai terminologisti arabi. Il procedimento di derivazione è la fonte principale della formazione delle parole in arabo: secondo Khusara il 95,5 % delle terminologie e delle parole in arabo viene formato tramite la derivazione[134].

Essenziale nel sistema derivazionale arabo è il concetto della radice che è un insieme di consonanti rappresentanti una nozione generale la quale acquisterà aspetti cognitivo- semantici diversi grazie all'aggiunta di altre lettere vocaliche, dette *pattern*[135]. Il numero delle radici in arabo è stimabile

133 Fahmi, H.H., *AlmrjE fY tEryb AlmSTlHAt AlElmyp wAlfnyp wAlhndsyp*, Il Cairo, AlnhDp AlmSryp, 1961, p.18

134 Khasarah, M.M, *Elm AlmSTlH wTrA}q wDE AlmSTlHAt fy AlErbyp*, Damasco, Dar El Fikr, 2008 .p.76

135 Ryding riporta la seguente definizione dei pattern in arabo: "A pattern is a bound and in many cases, discontinuous morpheme consisting in one or more vowels and slots for root phonemes (radicals), which

tra 5000 e 6500[136]. A seconda del numero delle consonanti all'interno di loro le radici si classificano in bilettere, trilettere, o quadrilettere, come dimostra l'esempio seguente:
- bt (tagliare) : bilettere
- ktb (scrivere) : trilettere
- dhrj (rotolare): quadrilettere

In arabo la derivazione è un metodo produttivo della formazione lessicale, in quanto da una sola radice si possono creare più di 400 forme lessicali diverse[137]. Le tipologie principali del procedimento di derivazione in arabo sono tre:

- **La derivazione morfologica**, detta anche *derivazione minore*. Questo tipo di derivazione crea, a partire da una radice verbale, nuove forme lessicali che condividono con l'originale radicale una relazione cognitivo-semantica e conservano anche le lettere principali della radice nonché il loro ordine all'interno della parola, come per es. dalla radice "ktb" (scrivere) derivano le parole seguenti:
- **ktb**: lui ha scritto
- **k**A**tb**: scrittore
- **kt**A**b**: libro
- **kt**A**b**p: scrittura
- m**kt**w**b**: scritto
- m**k**A**tb**p: corrispondenza
- m**ktb**p: biblioteca

2.3.2 La composizione: anche se non è tanto comune nella morfologia araba tradizionale, formare parole tramite la combinazione di due parole si adopera nell'arabo moderno standard per coniare soprattutto nuovi termini tecnici. Esempi di termini tecnici formati tramite la composizione sono *r>smAl* (il

---

either alone or in combination with one to three derivational affixes, interlocks with a root to form a stem, and which generally has grammatical meaning". Karin C. Ryding, *A reference Grammar of Modern Standard Arabic*, Cambridge University Press, 2005, p.48

136 Ibidem

137 Fahmi, H. H., *AlmrjE fY tEryb AlmSTlHAt AlElmyp wAlfnyp wAlhndsyp*, op cit, , p.20

capitale), formato dalle due parole *r>s* (testa) e *mAl* (soldi); *lAmrkzyp* (decentralizzazione) composta da *lA* (non) e *mrkzyp* (centralizzazione). In alcuni casi i termini creati, per la maggior parte in risposta a necessità di traduzione da lingue straniere sotto forma di calchi, consistono in due parole, formando un sintagma nominale, come <nEdAm wjwd (non-esistenza); mtEdd Al>TrAf (multilaterale)[138].

2.3.3 La sostituzione: è un procedimento di coniare nuove parole tramite la sostituzione di qualche lettera con un'altra all'interno di una parola già esistente per motivi sia morfologici che lessicali[139]. Esempi dei derivati per sostituzione:
- txdyr / txtyr (anestesia)
- mxAT/mgAT (muco)

2.3.4 La terminologizzazione: è un procedimento per formare nuovi tecnicismi in un dominio scientifico tramite uno di questi modi:

- Prestito da altre lingue o arabizzazione: non sono poche le parole provenienti da lingue straniere (soprattutto dall'inglese) nella lingua araba e in particolare nei linguaggi specializzati. In maggiori casi questi termini di provenienza straniera vengono arabizzati, adattandosi cioè alle regole fonetiche e morfologiche della lingua araba.

Arabicization has also has served Arabic as one of the most practical method of creating Arabic neologisms and terminology since the beginning of the nineteenth century when the role of Arabic as a transmitter language began to decline. Arabicization is more effective in handling new technical and scientific terms than both derivation and blending[140].

138 Karin C. Ryding, op cit., p.50
139 Khasarah, M.M, op cit., p.148
140 Elmgrab, R.A., "Methods of Creating and Introducing New Terms in Arabic: contributions from English-Arabic Translation". In *International Conference on Languages, Literature and Linguistics*, IPEDR vol.26 (2011), Singapore, IACSIT Press, p.499

Anche se hanno corrispondenti in arabo, molti prestiti stranieri sono diventati praticamente termini arabi solidi, come per es.

- btrwl (petrolio)
- kmbywtr (computer)
- tlyfwn (telefono)
- tlfzywn (televisone)
- bnk (banca)
- fylm (film)
- fydyw (video)
- kmbyAlp (cambiale)
- bwlySp (polizza)

- Prestito interno di tecnicismi: è un prestito terminologico in cui un tecnicismo in un dominio scientifico passa a un altro linguaggio specializzato tramite una modificazione semantica, come per es[141].

- xly (alveare (apicoltura)) → xlyAt (cellule (biologia))

- mHTAt (stazione (telecomunicazione))→ mHTAt (stazione (scienza spaziale))

141 Reguigui, A, op.cit, p.81

Capitolo III: Il corpus della tesi

3.1. I corpora linguistici

Con il crescente sviluppo delle tecnologie informatiche che consentono di raccogliere, gestire ed esplorare enormi quantità di dati linguistici, l'interesse alla creazione di corpora linguistici[142] è cresciuto recentemente in una maniera esponenziale. È indubbio che oggigiorno l'enorme disponibilità dei dati sul web ha agevolato significativamente la costituzione e la distribuzione dei corpora linguistici sia i corpora monolingui che quelli multilingui.

L'espansione delle capacità di memorizzazione ed esplorazione dei dati da parte dei computer, lo sviluppo di metodi avanzati per il trattamento e la codifica dei testi digitali e la disponibilità crescente di materiale testuale già in formato digitale hanno permesso non solo un ampliamento quantitativo dei corpora, ma anche una forte evoluzione qualitativa, che concerne i tipi di testi che vengono a comporre il corpus, il modo in cui questi vengono rappresentati in formato digitale e anche gli strumenti per la loro esplorazione[143].

Mentre originalmente il termine *corpus* indicava un insieme di testi, molto spesso di un unico autore, in formato annotato o meno, lo stesso termine si è evoluto recentemente per comprendere, come sostiene Baker[144], tre ulteriori caratteristiche fondamentali quali sono:
- un corpus deve avere un formato elettronico e essere analizzabile automaticamente o semi-automaticamente;
- un corpus non deve contenere solo testi scritti, bensì può comprendere altresì registrazioni o discorsi orali;
- un corpus può racchiudere generi testuali provenienti da differenti fonti con varietà di temi o di autori.

142 Secondo la definizione di Barbera i corpora linguistici sono "Raccolta di testi (scritti, orali o multimediali) o parti di essi in numero finito in formato elettronico trattati in modo uniforme (ossia tokenizzati ed addizionati di markup adeguato) così da essere gestibili ed interrogabili informaticamente; se (come spesso) le finalità sono linguistiche (descrizione di lingue naturali o loro varietà), i testi sono perlopiù scelti in modo da essere autentici e rappresentativi,": Barbera, M. *Linguistica dei corpora e linguistica dei corpora italiana. Un'introduzione*, Milano, Qu.A.S.A.R.s.r.l, 2013, p,18

143 Lenci, L., et al, *Testo e computer: elementi di linguistica computazionale*, Roma, Carocci editore, 2012, p.34

144 Baker, M., "Corpora in Translation Studies. An Overview and Suggestions for Future Research". In *Target*, 7(2). 1995. p. 225

Le prime applicazioni della linguistica dei corpora erano indirizzate, secondo Candin[145], principalmente verso gli studi linguistici, e in particolare quelli lessicografici. Successivamente e soprattutto dopo lo sviluppo avvenuto nella tecnologia informatica nonché l'affermarsi degli approcci socio-comunicativi, discorsivi, testuali, ecc., che invitano a studiare i fenomeni linguistici a partire dal loro ambiente contestuale, i corpora linguistici si sono evoluti per costituire una risorsa essenziale per altri campi di ricerca linguistica come la didattica delle lingue straniere, l'analisi del discorso, la traduzione automatica, ecc.[146], in quanto "le generalizzazioni o le informazioni statistiche sulle lingue derivate da corpora che rappresentano una gamma di registri tendono a fare la media dei risultati e possono quindi mascherare importanti differenze tra le varietà delle lingue"[147].

L'interesse alla costituzione dei corpora, soprattutto quelli specializzati, è stato affiancato effettivamente dal progresso nel campo della traduzione automatica e della terminologia computazionale, i cui studi risultano legati fortemente ai corpora.

Corpus-based research has become widely accepted as a factor in improving the performance of machine translation systems, and corpus-based terminology compilation is now the norm rather than the exception[148].

Oltre alla traduzione automatica, nell'ambito della linguistica computazionale i corpora linguistici, e in particolare quelli paralleli, acquistano un'importanza assoluta, soprattutto per applicazioni come l'estrazione di terminologie o la disambiguazione semantica. Inoltre, i corpora rappresentano una fonte essenziale per l'addestramento e la valutazione degli algoritmi informatici. In effetti, mentre i corpora linguistici "forniscono la base per l'estrazione delle regolarità di frequenza e co-occorrenza che servono

145 Gandin, S., "Linguistica dei corpora e traduzione: definizioni, criteri di compilazione e implicazioni di ricerca dei corpora paralleli". In *AnnalSS 5*, 2005 (2009), p.134
146 Cfr. Zanettin, F., *Translation-driven corpora: Corpus resources for descriptive and applied translation studies*. Manchester: St. Jerome Publishing, 2012, p.8
147 Kennedy, G., *An Introduction to Corpus Linguistics*, London, New York, Longman, 1998, p.279
148 Baker, M., "Corpora in Translation Studies" op cit., p. 223

per elaborare gli algoritmi di analisi degli applicativi probabilistici", sono proprio questi applicativi probabilistici che si utilizzano per creare e annotare i corpora che poi vengono adoperati per affinare i modelli e gli algoritmi e così via tramite un processo di interdipendenza o "circolo virtuoso" come lo definisce Chiari[149].

3.2. Tipologie dei corpora linguistici

I criteri fondamentali in base ai quali i corpora linguistici vengono designati o classificati si possono riassumere in:
1. generalità: in questo senso possiamo distinguere tra corpora che appartengono alla lingua comune e corpora provenienti da linguaggi settoriali o specialistici. Mentre nel primo caso si possono trovare rappresentati in un unico corpus vari registri linguistici che hanno come origine differenti contesti comunicativi, nell'altro caso i corpora si dimostrano centrati su un specifico genere testuale oppure su una ristretta varietà linguistica con delle caratterizzanti particolarità linguistiche, come per es. il dominio giuridico, tecnico, medico, sportivo, ecc.;
2. modalità: questo parametro classifica i corpora in corpora di testi scritti vs. corpora di lingua parlata[150];
3. cronologia: in base all'elemento cronologico i corpora possono essere sincronici o diacronici;
4. lingua: a seconda della lingua i corpora si possono etichettare in corpora monolingui e corpora bilingui o multilingui.

I corpora multilingui si dividono a loro volta in corpora paralleli e corpora comparabili.

Un *corpus parallelo* si può definire come "un corpus formato da una serie di testi originali in una determinata *lingua di origine* (definita tecnicamente

149 Chiari, I., "La chiave probabilistica delle lingue: teoria linguistica e applicazioni computazionali", in L. Fulci e E. Sciubba (a cura di), *Linguaggio, Mente e Società*, Roma, EuRoma-La Goliardica, 2008, p.75
150 Cfr. Sinclair, J., *Corpus, Concordance, Collocation*, Oxford, Oxford University Press, 1991, p.15

anche *Source Language*, o SL) e dalle relative traduzioni in un altra *lingua* (o altre lingue) *di destinazione* (*Target Language*, o TL)"[151]. Un *corpus comparabile* è invece una collezione di testi dello stesso dominio o genere testuale, creati in un'unica lingua (corpora comparabili monolingui), oppure in più lingue (corpora comparabili multilingui)[152].

Fra le tipologie dei corpora appena esposte i corpora paralleli rappresentano un'importanza di rilievo in particolar modo per gli studi di traduzione in quanto "forniscono un valido supporto per effettuare analisi linguistiche approfondite sui più importanti aspetti lessico-grammaticali e stilistici che contraddistinguono le metodologie traduttive di determinati generi linguistici, per poterne così individuare le caratteristiche di utilizzo, i punti di forza, i limiti e gli eventuali aspetti da migliorare"[153].

Per creare i corpora paralleli, di cui fa parte il nostro corpus, ci sono tanti modelli e combinazioni da adottare e che Candin espone nei seguenti schemi:

- il *modello uni-direzionale*: è un corpus parallelo che comprende testi in una sola lingua di partenza e corrispettive traduzioni in una sola lingua d'arrivo;
- il *modello bi-direzionale*[154], che è un corpus che contiene testi scritti in due lingue originali insieme con le relative traduzioni nelle medesime due lingue;
- il *modello a stella*[155]: si tratta di un corpus la cui lingua di partenza è unica mentre le lingue d'arrivo sono più di una;

151 Gandin, S., Linguistica dei corpora e traduzione, op cit, p.134
152 Olohan, M., "Introducing Corpora". In *Translation Studies*, London&New York: Routledge, 2004, p.35. Fantinuoli &Zanettin sostengono che la distinzione tra le due definizioni non è netta, dal momento che da un lato ci sono corpora multilingui definiti paralleli, che contengono solo testi originali, come per es. *Europarl*, e, dall'altro, i corpora comparabili possono avere vari gradi di similarità e comprendere non solo testi originali, ma anche traduzioni: "It may thus be useful to consider the attribute "parallel" or "comparable" as referring to a type of corpus architecture, rather than to the status of the texts as concerns translation". Fantinuoli, C. & Zanettin, F., "Creating and using multilingual corpora in translation studies". In *New directions in corpus-based translation studies*, Berlin:Language Science Press, 2015 p.4
153 Gandin, S., Linguistica dei corpora e traduzione, op cit, p.143
154 Cfr. Johansson, S.,, "Reflection on Corpora and their Uses in Cross-linguistic Research" in Zanettin, F., Bernardini S., Stewart D.,(a cura di) *Corpora in Translator*, Manchster, St. Jerome: 2003, p.138
155 Iva, p.140

- il *modello a diamante*[156]: si tratta di un modello abbastanza complicato per il numero delle combinazioni presenti, in quanto quel modello prevede testi originali creati in tre o più lingue (per es. testi originali scritti in italiano, francese, inglese) e le corrispettive traduzioni combinate (es. traduzioni dall'italiano all'inglese e al francese, dal francese all'italiano e all'inglese e dall'inglese all'italiano e al francese).

3.3. Disegno dei corpora linguistici

I criteri fondamentali che vanno presi in considerazione al momento della creazione dei corpora comprendono i seguenti aspetti che variano di importanza a seconda della tipologia nonché degli obiettivi del corpus:

- il dominio o l'area tematica del corpus[157]. La determinazione del campo tematico di un corpus si considera fra le iniziali fasi che condizionano la sua compilazione e questo per il fatto che "delineare chiaramente l'area tematica di ricerca consente di stabilire i criteri di selezione dei testi, le modalità di ricerca e gli obbiettivi del progetto, determinando di conseguenza tutte le scelte relative alla creazione del corpus.[158]"

- La dimensione del corpus. Insieme alla rappresentatività la questione di dimensione incide sulla "validità e affidabilità" di un corpus, anche se il concetto della grandezza rimane ancora non ben definito dal momento che qualsiasi corpus, quantunque sia la sua dimensione, non può essere che un piccolo campione del tipo linguistico che rappresenta[159]. La dimensione del corpus non appare, tuttavia, una decisione che viene presa deliberatamente da chi lo costituisce, bensì è una scelta vincolata ad un insieme di fattori, come la disponibilità in misura quantitativamente e qualitativamente appropriata dei testi, gli obiettivi prestabiliti della creazione del corpus, il livello finanziario del progetto (cioè se la costituzione del corpus viene realizzata da individui,

156 Iva, p.139
157 Gandin, S., Linguistica dei corpora e traduzione, op cit, p.143
158 Ibidem
159 Kennedy, G., *An Introduction to Corpus Linguistics*, op cit, p.66

istituti di ricerca, governi, ecc.), nonché lo spazio di tempo dedicato all'attuazione del progetto[160].

Size often becomes a decisive factor in assessing the reliability of corpus findings, since statistical considerations play an important role when it comes to drawing generalizations about one state of affairs in a language or language variety, for instance, when defining the meaning and usage of a word or making hypotheses about universal properties of translation[161].

- La rappresentatività del corpus: la rappresentatività "agisce come vincolo qualitativo e quantitativo sulla capacità del corpus di fornirci un modello in scala delle proprietà di una lingua o di una sua varietà"[162]. Tuttavia, per poter sapere fino a che punto un campione può rappresentare lo spazio della variabilità inclusa nella popolazione, e quindi valutare effettivamente la rappresentatività del campione, bisogna innanzitutto, come sostiene Biber[163], dare una precisa definizione della popolazione che il campione dovrebbe rappresentare. Secondo Biber tale definizione ha due aspetti: definire i limiti della popolazione, cioè individuare quali testi possono fare parte del corpus e quali invece saranno esclusi; e l'organizzazione categorica dei testi all'interno della popolazione.

Tuttavia, visto il numero delle variabili da prendere in considerazione al momento della creazione del corpus, che potrebbe non consentire in tutti i casi di costituire il corpus "ideale", il concetto di rappresentatività appare, come afferma Zanettin, molto sfuggente, spesso qualcosa per cui si deve lottare piuttosto che qualcosa che ragionevolmente essere raggiunto[164].

- La qualità dei testi selezionati: un elemento fondamentale che va preso in considerazione al momento della selezione dei testi da inserire nel corpus è la

160 Bowker, L., J. Pearson, J., *Working with Specialized Languages: A practical guide to using corpora*, London&New York: Routledge, 2002, p.45
161 Zanettin, F., "Translation-driven corpora", op cit, p.42:3
162 Lenci, A. et al., *Testo e Computer*, op cit, p.36
163 Biber, D., "Representativeness in Corpus Design". In *Literary and Linguistic Computing*, Volume 8, n.4, 1993, p.243
164 Zanettin, F., "Translation-driven corpora", op cit, p.46

misura qualitativa dei testi, che comprende a) la dinamicità o la staticità dei documenti, cioè se si tratta di un corpus aperto o chiuso in base alle prospettive legate allo scopo della progettazione del corpus; b) l'affidabilità o l'ufficialità delle fonti dei testi; c) la qualità di traduzione in caso di corpora paralleli; d) l'uso legale dei documenti il che significa l'autorizzazione da parte degli autori dei testi in caso che sia previsto il diritto d'autore sui documenti costituenti il corpus. Per quanto riguarda quest'ultimo punto bisogna dire che la questione del diritto d'autore nella creazione e distribuzione dei corpora appare assai "spinosa"[165] non perché le norme riguardanti la proprietà intellettuale variano da uno Stato all'altro, ma anche per la mancanza di una netta definizione del concetto di distribuzione dei corpora. Una variabile molto importante nel chiedere il permesso del diritto d'autore è determinare se l'utilizzo del corpus è per la ricerca scientifica oppure per motivi commerciali.

Malgrado l'evidente rilievo dei corpora paralleli sia per gli studi di traduzione che per la ricerca linguistica, non tutte le lingue umane risultano partecipanti egualmente a questo tipo di corpus. In effetti, la lingua araba, per motivi riguardante la modesta disponibilità sul web di testi paralleli in arabo e in altre lingue nonché la complessità del sistema morfologico arabo, presenta una limitata partecipazione a corpora paralleli, soprattutto a quelli specialistici.

Tale mediocre presenza dei corpora arabi nello stato dell'arte dei corpora paralleli non corrisponde effettivamente all'importanza della lingua araba fra le lingue umane nel mondo soprattutto se si considera il parametro del numero dei parlanti l'arabo nel mondo: l'arabo è la lingua ufficiale di tutti i ventidue paesi che aderiscono alla Lega Araba, con quasi 237 milioni di persone che parlano l'arabo come prima lingua, il che lo mette nel quinto posto nella classifica delle prime 100 lingue parlate nel mondo ordinate per numero di madrelingua. Ciononostante si è cominciato solo recentemente a mostrare qualche interesse al trattamento automatico della lingua araba, particolarmente tramite ricerche e studi contrastivi con la lingua inglese.

165 Ivi, p.52

Interest in Arabic is increasing as it gains importance for political, strategic, and business reasons. This interest has given rise to projects targeted at developing fast and accurate Arabic-English Machine Translation (MT). These efforts have been mostly spearheaded by the U.S. Department of Defense, and industrial leaders such as Google or IBM, and a few smaller companies in Europe, United States and Egypt.[166]

Tuttavia, si tratta di un interesse non privo di qualche sfida viste le particolari caratteristiche linguistiche della lingua araba. In effetti, l'accuratezza dei risultati del TAL, nato con l'obiettivo di trovare le tecniche che possano capire e analizzare il linguaggio umano, risente ovviamente della struttura linguistica della lingua di cui ci si occupa, in quanto la ricchezza morfologica e la complessità sintattica di una lingua, come è il caso dell'arabo, influiscono sullo sviluppare dei sistemi efficaci in grado di analizzare e interrogare automaticamente i testi. Delle peculiarità linguistiche dell'arabo si può ricordare: a) l'arabo ha un sistema morfologico evidentemente complesso, caratterizzato da una struttura flessiva e pronominale molto ricca; b) in arabo non si usano le lettere maiuscole; d) le forme verbali e nominali cambiano forma a seconda della posizione nella frase per effetto, cioè, dei casi di declinazione; e) l'arabo standard che è una forma semplificata dell'arabo classico e che viene usato nella maggior parte dei campi della vita non porta l'uso delle vocali brevi, cioè quei segni grafici che corrispondono ai diversi suoni e che si utilizzano per disambiguare i significati delle parole, ecc..

166 Zbib, R., Soudi, A., "Introduction: Challenges for Arabic Machine Translation". In Soudi, A. et al.(a cura di), *Challenges for Arabic Machine Translation*, John Benjamins Publishing Company, 2012, p.1

3.4. Stato dell'arte dei corpora italiani e arabi

A nostra conoscenza, l'unico corpus parallelo italiano-arabo è *L'arabo per la 488*[167]che comprende un insieme di testi generici: si tratta di progetto finalizzato allo sviluppo di strumenti e risorse tanto per la lingua italiana quanto per la lingua araba con particolare attenzione all'aspetto contrastivo.

Per quanto concerne, invece, lo stato dell'arte delle nostre due lingue come partecipi insieme ad altre lingue di corpora paralleli, troviamo che l'italiano prende parte a risorse testuali multilingue in misura maggiore rispetto all'arabo.

La maggior parte delle risorse linguistiche cui prende parte l'arabo è stata creata dal Linguistic Data Consortium (LDC)[168], nato con l'obiettivo di creare e distribuire risorse linguistiche necessarie per lo sviluppo e l'addestramento di strumenti computazionali in grado di capire e analizzare le lingue umane. LDC comprende una grande varietà dei documenti arabi, raccolti soprattutto da giornali, gruppi di notizie, blog, email, ecc.. Accanto ai corpora monolingui, LDC possiede anche corpora paralleli arabo-inglese, specialmente nel dominio giornalistico, come per es.:

- GALE Phase 1 Arabic Broadcast News Parallel Text
- GALE Phase 1 Arabic Blog Parallel Text
- GALE Phase 1 Arabic Newsgroup Parallel Text
- ISI Arabic-English Automatically Extracted Parallel Text
- Multiple-Translation Arabic (MTA) Parts 1 and 2

Per quanto concerne i corpora paralleli in arabo e altre lingue si rammenta, inoltre, EAPCOUNT[169]. Si tratta di un corpus parallelo inglese-arabo con 341

167 Picchi E. et al., "Risorse monolingui e multilingui. Corpus bilingue italiano-arabo". In *Linguistica computazionale*, XVIII/XIX, 1999, Pisa

168 https://www.ldc.upenn.edu

169 Hammouda S, "Small Parallel Corpora in an English-Arabic Translation Classroom: No Need to Reinvent the Wheel in the Era of Globalization". In Shiyab, S., et al.(a cura di), *Globalization and Aspects of Translation*, UK: Cambridge Scholars Publishing, 2010

documenti delle Nazioni Unite allineati a livello di paragrafo. Poi, si menziona il corpus parallelo multilingue (inglese- spagnolo- arabo) creato presso il laboratorio di linguistica computazionale dell'università autonoma di Madrid[170]. Anche questo corpus contiene una collezione dei documenti delle Nazioni Unite, allineati a livello di frase e annotati morfosintatticamente.

Dei corpora paralleli in italiano e altre lingue ricordiamo *Bononia Legal Corpus*[171]. È un corpus inglese-italiano di testi giuridici paralleli e comparabili, sviluppato presso l'università di Bologna. Il progetto è stato creato in due fasi: nella prima si è costruito un corpus pilota, composto da corpora paralleli inglese-italiano; mentre nella fase successiva vengono aggiunti corpora comparabili, sempre nelle medesime lingue, riguardanti testi nell'ambito legislativo, giudiziario e amministrativo per analizzare le caratteristiche linguistiche dei due sistemi legali.
Inoltre, nell'ambito del progetto CATEX (*Computer Assisted Terminology Extraction*) presso l'Accademia Europea di Bolzano è stato realizzato un corpus giuridico parallelo italiano-tedesco[172]. Questo corpus, allineato a livello di frase, comprende una raccolta di leggi italiane con la relativa traduzione in tedesco con una dimensione di quasi 5 milioni di parole.

3.5. Descrizione del corpus della tesi

Non c'è dubbio che la progettazione di un corpus linguistico appare condizionata in primo luogo dagli obiettivi della ricerca che si intende effettuare nonché dalla qualità e dalla disponibilità dei testi.

Come dominio tematico del corpus abbiamo scelto il diritto internazionale e in particolare i diritti umani nel mondo. La scelta di questo genere testuale ha le seguenti motivazioni:

---

170 Samy, D., et al., "Building a Multilingual Parallel Corpus Arabic-Spanish-English". In *Proceedings of International Conference on Language Resources and Evaluation LREC-06*, 2006, Genoa, Italy
171 Favretti R. et al., "Words from Bononia Legal Corpus". In *Text Corpora and Multilingual Lexicography*, John Benjamins Publishing Company, 2007
172 Gamper, J., "CATEX– A Project Proposal". In *Academia*, 14, 10-12, 1998

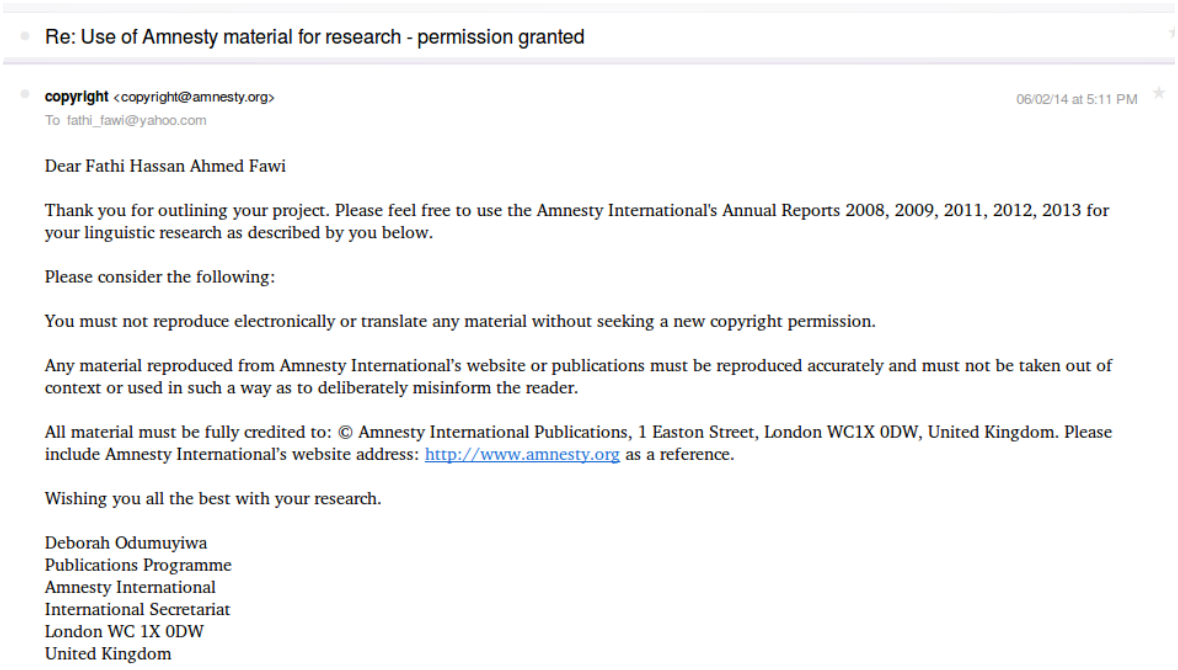
- Il linguaggio giuridico è uno dei linguaggi specializzati che presentano molte peculiarità a diversi livelli dell'analisi linguistica, il che rende indifferibilmente necessario fornire e sviluppare corpora di testi giuridici, soprattutto tra due lingue che appartengono a due famiglie linguistiche diverse come l'italiano e l'arabo;

- Per quanto riguarda la lingua araba, la maggior parte dei corpora giuridici disponibili sul web riguarda il codice di famiglia degli Stati arabi, che, ispirato ai principi della *Shariah* islamica, contiene tante terminologie islamiche che non hanno corrispondenti in italiano. Per questo problema dell'intraducibilità dei termini giuridici islamici, abbiamo pensato quindi al diritto internazionale, dove risulta limitata l'influenza della dimensione religiosa sui termini;

- L'accuratezza della traduzione dei testi paralleli è un fattore essenziale soprattutto trattandosi di terminologie giuridiche, e nei documenti di Amnesty International (AI) e dell'Organizzazione delle Nazioni Unite (ONU) abbiamo trovato un livello di traduzione tanto accurato;

- I testi del corpus sono disponibili sul web e scaricabili facilmente in formato PDF;

- I diritti d'autore sono stati rispettati nell'ambito dell'uso legale dei testi. Quest'uso legale del nostro corpus ha due aspetti: a) i documenti dell'ONU non hanno diritto d'autore come è specificato dallo statuto dell'ONU stesso; b) per i testi dell'AI ci è stato fornito, su una nostra richiesta, un permesso scritto da parte dei detentori del diritto d'autore, che ci autorizza a utilizzare i loro documenti per i fini della presente ricerca. Riportiamo di sotto il testo di questa autorizzazione:



Re: Use of Amnesty material for research - permission granted

copyright <copyright@amnesty.org> 06/02/14 at 5:11 PM
To fathi_fawi@yahoo.com

Dear Fathi Hassan Ahmed Fawi

Thank you for outlining your project. Please feel free to use the Amnesty International's Annual Reports 2008, 2009, 2011, 2012, 2013 for your linguistic research as described by you below.

Please consider the following:

You must not reproduce electronically or translate any material without seeking a new copyright permission.

Any material reproduced from Amnesty International's website or publications must be reproduced accurately and must not be taken out of context or used in such a way as to deliberately misinform the reader.

All material must be fully credited to: © Amnesty International Publications, 1 Easton Street, London WC1X 0DW, United Kingdom. Please include Amnesty International's website address: http://www.amnesty.org as a reference.

Wishing you all the best with your research.

Deborah Odumuyiwa
Publications Programme
Amnesty International
International Secretariat
London WC 1X 0DW
United Kingdom

- Rappresentatività del corpus

La scelta dei testi che compongono il campione dovrebbe essere il più accurata possibile in modo che il corpus possa permettere di generalizzare le proprietà linguistiche del corpus all'intera popolazione. Tuttavia, se nei corpora generali la rappresentatività appare abbastanza complessa, vista la variabilità enorme all'interno della collezione dei testi, nei corpora specialistici la questione sembra, invece, abbastanza controllabile, in quanto in questo caso i testi scelti appartengono ad un unico registro linguistico. Come è stato detto precedentemente, il nostro corpus contiene testi che appartengono al diritto internazionale. Per quanto riguarda i documenti dell'ONU, i testi risultano rappresentativi del linguaggio giuridico nelle due lingue a livello sia sintattico che lessicale. Per i documenti dell'AI si notano, invece, due caratteristiche inerenti alla rappresentatività del corpus: da una parte i testi hanno uno stile sintattico vicino a quello della lingua comune, dall'altra ci si nota la presenza notevole di termini presi da altri domini,

sopratutto da quello politico. Nel caso dei testi dell'AI la questione riguardante le strutture sintattiche non rappresenta un problema nel nostro caso perché il nostro studio si concentra solo sulle terminologie e quindi non si interessa alle caratteristiche sintattiche delle frasi. Per quanto riguarda la presenza di termini non giuridici nel corpus, il nostro procedimento provvede a selezionare dal corpus parallelo solo i termini giuridici.

I testi del corpus si dividono in due categorie: i documenti dell'AI e una varia raccolta di convenzioni internazionali nel campo del diritto internazionale. L'AI è un'organizzazione non governativa internazionale impegnata nella difesa dei diritti umani in tutto il mondo. Lo scopo dell'organizzazione è quello di promuovere, in modo indipendente e imparziale, il rispetto dei diritti umani enucleati nella Dichiarazione universale dei diritti umani e negli altri standard internazionali relativi ai diritti umani. Le sezioni dell'AI comprendono quasi tutte le lingue del mondo. Ogni anno l'organizzazione pubblica un rapporto in diverse lingue, in cui analizza la situazione dei diritti umani nel mondo. Di questi rapporti abbiamo scelto cinque rapporti annuali (2008, 2009, 2011, 2012, 2013). Ai documenti dell'AI abbiamo pensato di aggiungere altri generi testuali per garantire una certa rappresentatività dei termini. Ci siamo serviti, quindi, dei documenti dell'ONU. Si tratta di una grande raccolta di accordi, convenzioni, protocolli internazionali sempre nell'ambito del diritto internazionale in generale e dei diritti umani in particolare. Quest'ultima parte del corpus comprende due categorie testuali: la prima racchiude un insieme di convenzioni e accordi internazionali nell'ambito dei diritti umani nel mondo, mentre la seconda contiene le convenzioni dell'Organizzazione Internazionale del Lavoro (ILO).

In totale il corpus comprende all'incirca 2,7 milioni di parole.

	n. token	n.frasi	lunghezza media delle frasi	type/token ratio
Italiano	1488099	64482	30	0.043
Arabo	1272726	61922	39	0.077

Tabella (1): statistica del corpus della tesi

3.6. Costituzione e preparazione del corpus

3.6.1. Raccolta e conversione dei testi

Il web rappresenta la fonte principale per la raccolta dei dati del corpus, sia per i testi italiani che per quelli arabi. Il risultato di questa fase è una raccolta di documenti in formato PDF in entrambe le lingue. Il formato PDF non consente, tuttavia, un trattamento automatico dei testi, quindi bisognava convertire i testi nel formato "Plain text format" che è adeguato a qualsiasi trattamento computazionale del corpus, e poi salvare i testi in UNICODE che è adeguato nel nostro caso dato che i sistemi di scrittura delle due lingue sono diversi.

Il processo della conversione non è, tuttavia, banale come sembra, soprattutto per la lingua araba. Fra le notevoli osservazioni individuate durante la conversione dei testi arabi ricordiamo: la perdita di alcuni caratteri, lo scambio tra certi caratteri (soprattutto tra "A" e "l"), l'inversione della direzione di scrittura (soprattutto i numeri), la perdita del formato del testo originale, ecc.. Tutto questo richiede un grande sforzo per rimuovere ogni forma di "rumore" e restituire la normalità dei testi. Nel caso dei testi italiani gli errori derivati dalla conversione riguardano maggiormente il cambiamento del formato del testo originale.

3.6.2. Trattamento del corpus

Fino al passo precedente lo stato del corpus è grezzo, cioè senza nessun'annotazione linguistica, che è utile per esplorare ed interrogare il corpus in modo migliore. Zanettin[173] distingue tra due tipi di annotazione dei corpora: a) annotazione procedurale o di presentazione, che concerne informazioni sulla formazione visiva di un testo, come per es. il tipo o la dimensione dei caratteri, dei titoli, ecc.; b) annotazione descrittiva o strutturale che riguarda, invece, il contenuto, sia linguistico che extra-linguistico, dei documenti componenti il corpus.

173 Zanettin, F., "Translation-driven corpora", op cit, p.74

L'importanza dei corpora annotati consiste non solo nella possibilità di esplorare ed estrarre informazioni dal testo, ma anche nel fornire "training e valutazione di algoritmi specifici in sistemi automatici"[174].

Il trattamento automatico del corpus passa per le seguenti fasi:

3.6.2.1 Segmentazione

La segmentazione dei testi è stata effettuata a livello di frase. Per segmentare i testi abbiamo utilizzato un algoritmo nel pacchetto NLTK[175] basato sulla punteggiatura (".", "?", "!"). Tuttavia, non mancano gli errori anche in questa fase; soprattutto per la mancanza dell'uso delle lettere maiuscole in arabo. Vista la natura giuridica dei testi, si sono registrate alcune peculiarità riguardanti i confini di frase nei testi del corpus. In questo caso il segno della fine frase non è solo il punto finale come è il caso dei testi generali, ma i segni ":", ";" si possono ritenere anche confine di frase, soprattutto quando iniziano una lista di clausole o commi. Il risultato di questa fase è un testo segmentato a livello di una sola frase per riga a livello monolingue.

3.6.2.2 Tokenizzazione

Tokenizzare un testo significa ridurlo nelle sue unità ortografiche minime, dette *token*, che sono unità di base per ogni successivo livello di trattamento automatico. La complessità di questo compito dipende in misura relativamente maggiore dalla natura della lingua del corpus. Mentre nelle lingue indoeuropee tokenizzare un testo appare semplice, in altre lingue come l'arabo la stessa funzione appare molto complicata considerando il suo sistema morfologico molto ricco. In arabo un'unica parola può presentare più di una soluzione morfologica, il che significa che per disambiguare al meglio le unità lessicali di un testo arabo ogni sistema di tokenizzazione necessita di un analizzatore morfologico.

---

174 Zotti, P., "Costruire un corpus parallelo Giapponese-Italiano. Metodologie di compilazione e applicazioni". In Casari, M., Scrolavezza, P. (a cura di), *Giappone, storie plurali*, Bologna, I libri di Emil-Odoya Edizioni, 2013

175 http://www.nltk.org

Per tokenizzare il corpus arabo abbiamo utilizzato il software MADA TOKAN[176]. Si tratta di un pacchetto per il trattamento automatico dell'arabo moderno standard, e in particolare per il tagging morfologico (disambiguazione), diacritizzazione, e tokenizzazione. La scelta di questo sistema di tokenizzazione deriva proprio dalla sua dipendenza da un analizzatore morfologico per disambiguare le parole del testo. Il sistema utilizza l'analizzatore morfologico AL-MORGEANA[177] che fornisce per ogni parola nel testo le sue possibili analisi che vengono poi convalidate, tramite diversi modelli probabilistici, dal sistema MADA per determinarne, considerando anche il contesto, l'analisi più giusta che rappresenta, infine, l'input del sistema di tokenizzazione TOKAN.

Un ulteriore vantaggio di questo sistema è la possibilità di controllare l'output del testo tokenizzato tramite un file di configurazione, detto TOKAN_SCHEME, che comprende quattro variabili controllabili dall'utente: *Single Variables* che stabilisce il formato dell'intero schema di tokenizzazione, come per es. il tipo di encoding dei testi, ecc.; *SPLIT Variables* che determina quali unità lessicali (articoli, congiunzioni, pronomi clitici, ecc.) da separare dal tema della parola di input; *FORM Variables* che riguarda il formato dell'output del testo tokenizzato, come per es. se l'output delle parole tokenizzate sarà lemma, tema, o nel formato originale nel testo; e infine *Aliases* il quale raggruppa in maniera sintetica gli schemi di tokenizzazione più comuni per facilitare agli utenti, soprattutto i non esperti, il processo di tokenizzazione.

Visti gli obiettivi del nostro studio, abbiamo scelto un formato di tokenizzazione che si riassume nel seguente modo:

1. TOKAN_SCHEME = ::SPLIT QUES CONJ PART REST PRON ::FORM0 WORD ENCODE:UTF8

176 Habash, N., Rambow, O., "Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In *Proceedings of the Conference of American Association for Computational Linguistics* (ACL 05). Ann Arbor, Michigan. 2005.

177 Habash, N. "Arabic morphological representations for machine translation". In *Arabic Computational Morphology: Knowledge- based and Empirical Methods, Text, Speech, and Language Technology*. Kluwer/Springer, 2005

2. TOKAN_SCHEME = ::SPLIT QUES CONJ PART DART REST PRON ::FORM0 LEXEME ENCODE:UTF8

Il primo schema di tokenizzazione segmenta una parola separandone le particelle interrogative, le congiunzioni, le preposizioni ed i suffissi, e restituisce poi il resto nel formato WORD, cioè non LEMMA o STEM, nel sistema di ENCODE = UTF8; mentre l'altro schema aggiunge al primo la separazione dell'articolo determinativo "Al" e la variable LEXEME che restituisce le parole del testo, dopo la tokenizzazione, nella forma lemmatizzata che è una forma convenzionale delle parole senza le loro caratteristiche morfologiche. Tale modo di tokenizzare ci fornisce due testi: uno con le parole originali senza lemmatizzazione, mentre l'altro è lemmatizzato.

ttx* kl dwlp EDw، End <nfA* AltzAmAthA bmwjb AlAtfAqyp bAlqDA' ElY AlEml Aljbry >w Al<lzAmY، tdAbyr fEAlp lmnE w<zAlp Alljw' <lY AlEml Aljbry >w Al<lzAmY wltzwyd AlDHAyA bAlHmAyp wsbl AlwSwl <lY wsA}l AlAntSAf AlmnAsbp wAlfEAlp، mn qbyl AltEwyD، wlmEAqbp mrtkbY AlEml Aljbry >w Al<lzAmY.	Testo arabo
ttx* kl dwlp Edw ، End <nfA* AltzAmAt hA b mwjb AlAtfAqyp b AlqDA' ElY AlEml Aljbry >w Al<lzAmy ، tdAbyr fEAlp l mnE w <zAlp Alljw' <lY AlEml Aljbry >w Al<lzAmy w l tzwyd AlDHAyA b AlHmAyp w sbl AlwSwl <lY wsA}l AlAntSAf AlmnAsbp w AlfEAlp ، mn qbyl AltEwyD ، w l mEAqbp mrtkby AlEml Aljbry >w Al<lzAmy .	Testo tokenizzato in forma originale
Atx* kl dwlp EDw ، End <nfA* AltzAm hA b mwjb Al AtfAqy b Al qDA' ElY Al Eml Al jbry >w Al <lzAmY ، tdbyr fEAl l mnE w <zAlp Al ljw' <lY Al Eml Al jbry >w Al <lzAmy w l tzwyd Al DHyp b Al HmAyp w sbl Al wSwl <lY wsylp Al AntSAf Al mnAsb w Al fEAlp ، mn qbyl Al tEwyD ، w l mEAqbp mrtkb Al Eml Al jbry >w Al <lzAmy .	Testo tokenizzato in forma lemmatizzata

Tabella (2). Esempio del corpus arabo tokenizzato

La scelta di tale strategia di tokenizzazione risponde effettivamente al tipo di studio della nostra tesi: separare i pronomi clitici dal resto della parola ci appare utile per consentire di toglierli, in una fase successiva, dal corpus, in modo che possiamo estrarre termini candidati senza troppe varianti pronominali che non rappresentano importanza nel nostro caso; mentre fornire una forma lemmatizzata del testo ci aiuta ad avere delle statistiche rilevanti e significanti dei termini candidati a prescidere dalle varianti flessive, come vederemo nel capitolo successivo parlando del filtro statistico dei termini. Nel nostro caso il tasso di correttezza nel compito di tokenizzazione dei testi arabi è all'incirca 98%, e gli errori individuati riguardano maggiormente errori di battitura riscontrati nel testo originale.

Per tokenizzare i testi italiani abbiamo utilizzato il tokenizzatore contenuto nel pacchetto LinguA[178] che comprende quattro componenti del trattamento automatico dell'italiano e dell'inglese, e in particolare: segmentazione delle frasi (Sentence splitting), tokenizzazione, PoS tagging e lemmatizzazione, e un parser a dipendenza. Vista la semplicità del compito, la tokenizzazione dei testi italiani non richiede certi schemi come nel caso dell'arabo. Similmente al corpus arabo, anche per il corpus italiano dopo la fase di tokenizzazione disponiamo di due corpora: uno con le parole originali e l'altro in formato lemmatizzato.

178 http://www.italianlp.it

1	ogni	ogni	D	DI	num=s|gen=n	2	mod	_	_		
2	Membro	Membro	S	SP	_	4	subj	_	_		
3	deve	dovere	V	VM	num=s|per=3|mod=i|ten=p	4	modal	_	_		
4	prendere	prendere	V	V	mod=f	0	ROOT	_	_		
5	misure	misura	S	S	num=p|gen=f	4	obj	_	_		
6	efficaci	efficace	A	A	num=p|gen=n	5	mod	_	_		
7	per	per	E	E	_	6	arg	_	_		
8	prevenire	prevenire	V	V	mod=f	7	prep	_	_		
9	ed	e	C	CC	_	4	con	_	_		
10	eliminare	eliminare	V	V	mod=f	4	conj	_	_		
11	l'	il	R	RD	num=s|gen=n	12	det	_	_		
12	utilizzo	utilizzo	S	S	num=s|gen=m	10	obj	_	_		
13	del	di	E	EA	num=s|gen=m	12	comp	_	_		
14	lavoro	lavoro	S	S	num=s|gen=m	13	prep	_	_		
15	forzato	forzare	V	V	num=s|mod=p|gen=m	14	mod	_	_		
16	,	,	F	FF	_	17	punc	_	_		
17	per	per	E	E	_	4	comp	_	_		
18	assicurare	assicurare	V	V	mod=f	17	prep	_	_		
19	alle	al	E	EA	num=p|gen=f	18	comp	_	_		
20	vittime	vittima	S	S	num=p|gen=f	19	prep	_	_		
21	una	uno	R	RI	num=s|gen=f	22	det	_	_		
22	protezione	protezione	S	S	num=s|gen=f	18	obj	_	_		
23	e	e	C	CC	_	18	con	_	_		
24	l'	il	R	RD	num=s|gen=n	25	det	_	_		
25	accesso	accesso	S	S	num=s|gen=m	18	obj	_	_		
26	a	a	E	E	_	25	comp	_	_		
27	meccanismi	meccanismo	S	S	num=p|gen=m	26	prep	_	_		
28	di	di	E	E	_	27	comp	_	_		
29	ricorso	ricorso	S	S	num=s|gen=m	28	prep	_	_		
30	e	e	C	CC	_	28	con	_	_		
31	di	di	E	E	_	28	conj	_	_		
32	risarcimento	risarcimento	S	S	num=s|gen=m	31	prep	_	_		
33	adeguati	adeguare	V	V	num=p|mod=p|gen=m	32	mod	_	_		
34	e	e	C	CC	_	33	con	_	_		
35	efficaci	efficace	A	A	num=p|gen=n	33	conj	_	_		
36	,	,	F	FF	_	17	punc	_	_		
37	come	come	B	B	_	39	mod	_	_		
38	l'	il	R	RD	num=s|gen=n	39	det	_	_		
39	indennizzo	indennizzo	S	S	num=s|gen=m	18	obj	_	_		
40	,	,	F	FF	_	17	punc	_	_		
41	e	e	C	CC	_	4	con	_	_		
42	per	per	E	E	_	4	conj	_	_		
43	reprimere	reprimere	V	V	mod=f	42	prep	_	_		
44	i	il	R	RD	num=p|gen=m	45	det	_	_		
45	responsabili	responsabile	S	S	num=p|gen=n	43	obj	_	_		
46	del	di	E	EA	num=s|gen=m	45	comp	_	_		
47	lavoro	lavoro	S	S	num=s|gen=m	46	prep	_	_		
48	forzato	forzare	V	V	num=s|mod=p|gen=m	47	mod	_	_		
49	o	o	C	CC	_	48	dis	_	_		
50	obbligatorio	obbligatorio	A	A	num=s|gen=m	48	disj	_	_		
51	.	.	F	FS	_	4	punc	_	_		

Tabella (3). Esempio del corpus italiano analizzato con LinguA

3.6.2.3 Annotazione morfo-sintattica del corpus

Per l'annotazione o l'etichettatura linguistica di un corpus si intende associare alle porzioni del testo informazioni linguistiche in forma di etichetta (tag o mark-up), sia per rendere esplicito il contenuto del testo sia per ottenerne una conoscenza approfondita. L'annotazione può riguardare qualsiasi livello di analisi linguistica del testo (fonetica, morfologia, sintassi, semantica, pragmatica). Scegliere il livello di annotazione dipende maggiormente dal tipo e dagli obiettivi del corpus. Il tipo di annotazione più conosciuto è quello morfosintattico o il cosiddetto PoS (part-of-speech) tagging, che consiste nell'attribuire ad ogni parola nel testo la sua categoria grammaticale. L'accuratezza di un tagger si basa sulla sua capacità di risolvere ogni possibile ambiguità assegnando ad ogni parola un tag corretto. In questo caso ci si serve o delle evidenze contestuali della parola o dei calcoli probabilistici da corpora già annotati. Tra i principali approcci utilizzati nel tagging ci sono quelli basati su regole (rule-based) e quelli statistici. Nel primo caso si formulano manualmente regole affinché il tagger possa assegnare correttamente le parti del discorso a ogni parola, mentre nel secondo caso un programma estrae regole e generalizzazioni da testi già annotati per poter taggare poi testi nuovi. Questo ultimo metodo di tagging si definisce come apprendimento automatico supervisionato, dove i programmi dipendono sempre dalle probabilità statistiche delle parti del discorso (tags) incontrate nel corpus di addestramento, e le utilizzano per taggare testi nuovi. Il compito del PoS tagging possiede un'importanza particolare nel trattamento automatico del linguaggio, in quanto rappresenta il primo passo nell'annotazione automatica dei testi, il che significa che gli errori che si riscontrano durante questa fase potrebbero incidere sui successivi passi o analisi del TAL. Nel nostro caso, il PoS tagging ci serve per formulare dei pattern morfo-sintattici tramite cui possiamo estrarre inizialmente dei termini candidati dai testi monolingui, per poi selezionarne solo i più specifici e rilevanti.

- Annotazione morfo-sintattica dei testi arabi

Sviluppare strumenti di PoS tagging per i corpora arabi è uno dei temi più importanti del TAL arabo. Secondo Habash le difficoltà principali che riguardano il TAL arabo hanno le seguenti motivazioni: a) la morfologia araba si serve di un ricco sistema di pronomi clitici e di forme flessive, come per es. la parola *wsyktbwnhA* (e loro la scriveranno) si compone delle seguenti parti: *w* (congiunzione) + *s* (particella del tempo futuro) + *y-ktb-wn* (il verbo più i segni di accordo di persona, genere e numero) + *hA* (pronome di complemento oggetto); d) l'arabo moderno standard non porta l'uso delle vocali brevi che si utilizzano per disambiguare i significati delle parole, il che crea in molti casi fonti di ambiguità. Basti pensare che ogni parola dell'arabo moderno ha mediamente 12 analisi nell'analizzatore morfologico (SAMA)[179]; c) l'arabo ha un numero grande di dialetti diversi che deviano a quasi tutti i livelli linguistici dalla lingua araba moderna. Queste sfide e altre vengono trattate nell'ambito del TAL con tecniche e metodi diversi a seconda del compito oppure della tipologia testuale. Il PoS tagging dei testi arabi ha visto ultimamente un interesse ben notevole da parte dei ricercatori che l'hanno affrontato con approcci diversi. Tuttavia, lungi da essere veri sistemi disponibili agli utenti, la maggior parte di questi lavori rimane solo a livello di tentativi senza implementazione. A nostra conoscenza ci sono due sistemi ben conosciuti nell'ambito del trattamento automatico della lingua araba in generale e del PoS tagging in particole: Amira[180] e MADA[181].

Amira è un sistema di PoS tagging basato sull'apprendimento supervisionato senza dipendenza da informazioni morfologiche. La tecnologia di Amira dipende dalle macchine a vettori di supporto (support vector machine:SVM) per la classificazione di pattern. Questo sistema comprende tre moduli per il trattamento automatico della lingua araba: tokenizzazione, PoS tagging, e

179 Graff, D., et al., *Standard Arabic Morphological Analyzer (SAMA) Version 3.1*. Linguistic Data Consortium LDC2009E73, 2009

180 Diab, M., "Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, PoS tagging, and base phrase chunking". In *2nd International Conference on Arabic Language Resources and Tools*, Cairo, Egypt, 2009

181 Habash, N., Rambow, O., Arabic Tokenization, op cit.

base-phrase chunked. Il sistema Amira è stato addestrato su Arabic Penn Treebank ATB.

La tecnologia di MADA si basa, invece, sull'analisi morfologica del testo per determinare la giusta categoria grammaticale dell'unità lessicale di input. Grazie all'analizzatore morfologico che per ogni parola di input fornisce le diverse possibili analisi, il sistema MADA può possedere, in una fase iniziale, di ogni parola nel testo la relativa possibile interpretazione morfologica, a diversi livelli (PoS tagging, lemma, flessione, pronomi clitici, ecc.). In una fase successiva, il sistema si serve di un insieme di modelli- SVM e modelli di linguaggio N-gram – per assegnare a ogni parola nel contesto le sue probabili caratteristiche morfologiche: categoria grammaticale, lemma, genere, numero e persona. La fase finale comprende la componente valutativa che, confrontando le analisi fornite dall'analizzatore morfologico con le caratteristiche morfologiche esplorate nella fase precedente, determina il tipo di analisi più pertinente per ogni parola nel testo.

Recentemente i due sistemi si sono congiunti in un unico sistema, detto MADAMIRA[182]. La performance del PoS tagging nella versione ottimizzata non ha portato, tuttavia, miglioramenti, anzi, in cambio di ridurre il tempo di esecuzione del programma, che si configura come il motivo principale della versione unica, si è registrata una piccola percentuale di peggioramento nel PoS tagging dell'arabo moderno standard.

Vista la semplicità del tagset di Amira nei confronti di quello di MADA, abbiamo deciso di affidare il PoS tagging del nostro corpus arabo al tagger di Amira, il cui tagset ERST (extended reduced tag set) comprende quasi 75 tag per le caratteristiche morfologiche di genere, numero, caso, modo, e definizione. Tuttavia, per motivi di semplificazione, abbiamo normalizzato tutti i tag per i nomi, limitando le categorie tra NN (nome), DET_NN (nome definito), JJ (aggettivo) e DET_JJ (aggettivo definito), come dimostra il brano seguente.

182 Pasha, A., et al., "MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic". In *Language Resources and Evaluation Conference(LREC)*, Reykjavik 2014.

tErD/VBD_MS3 Edd/NN mn/IN AlmdAfEyn/DET_NN En/IN Hqwq/NN Al<nsAn/DET_NN w/CC AlmnZmAt/DET_NN AlmEnyp/DET_JJ b/IN Hqwq/NN Al<nsAn/DET_NN l/IN Altrhyb/DET_NN w/CC Althdyd/DET_NN b/IN Swrp/NN mtzAydp/JJ ،/PUNC wsT/NN mnAx/NN mn/IN Alqywd/DET_NN ElY/IN Hryp/NN AltEbyr/DET_NN ،/PUNC $hd/VBD_MS3 >yDA/NN AlHkm/DET_NN b/IN sjn/NN >Hd/NN AlSHfyyn/DET_NN l/IN Edp/NN >$hr/NN ./PUNC

w/CC wrdt/CC_VBD_FS3 >nbA'/NN En/IN HAlAt/NN mn/IN Al<jlA'/DET_NN Alqsry/DET_JJ w/CC AnthAkAt/NN Hqwq/NN Al<nsAn/DET_NN ElY/IN >ydy/NNS_MD Al$rTp/DET_NN ،/PUNC w/CC <n/RP kAn/VBD_MS3 *lk/DT_MS ElY/IN nTAq/NN >ql/JJ mn/IN mvyl/NN fy/IN AlsnwAt/DET_NN AlsAbqp/DET_JJ ./PUNC

w/CC >sfr/NNP tmrd/NN l/IN AlsjnA'/DET_NN fy/IN sjn/NN lwAndA/NNP Almrkzy/DET_JJ En/IN sqwT/NN qtlY/NN w/CC jrHY/NN ،/PUNC w/CC lkn/VBP >EdAd/NN AlDHAyA/DET_NN kAnt/VBD_FS3 mwDE/NN xlAf/NN ./PUNC

Tabella(4):Esempio del corpus arabo taggato con Amira

Nel caso del nostro corpus la percentuale di accuratezza di Amira tagger arriva all'incirca del 95%. La maggior parte degli errori individuati nel corso del PoS tagging si classifica in tre tipi: a) nomi taggati erroneamente come aggettivi; b) aggettivi taggati erroneamente come nomi; c) verbi taggati erroneamente come nomi o aggettivi. Visti gli obiettivi della nostra ricerca che si interessa esclusivamente alle costruzioni sintagmatiche nominali nel corpus, i primi due tipi di errore non creano problemi durante le fasi successive di estrazione, dal momento che sia gli aggettivi che i nomi fanno parte dei pattern morfo-sintattici utilizzati nell'estrazione. Inoltre, il terzo tipo risulta gestibile, perché oltre alla sua scarsa rappresentanza, quasi 0,5%, sarà validato dalle misure di associazione lessicale, che scartano certamente ogni elemento inserito erratamente nella lista dei termini candidati.

- Annotazione morfo-sintattica dei testi italiani

Per i testi italiani si è utilizzato Felice-POS-Tagger[183]. Felice-POS-Tagger è una combinazione di sei tagger, con tre algoritmi diversi. Ognuno dei tre algoritmi viene utilizzato per costruire un left-to-right (LR) tagger e un right-to-left (RL) tagger. L'accuratezza del Felice-POS-Tagger nel taggare i testi del nostro corpus è all'incirca del 97%.

Difensori/S dei/EA diritti/S umani/A e/CC organizzazioni/S sono/VA stati/V oggetto/S di/E crescenti/S intimidazioni/S e/CC minacce/S in/E un/RI clima/S di/E limitazione/S della/EA libertà/S di/E espressione/S ,/FF che/PR ha/V anche/B visto/V un/RI giornalista/S incarcerato/V per/E diversi/DI mesi/S ./FS
Sono/VA stati/VA riportati/V casi/S di/E sgomberi/S forzati/A e/CC violazioni/S dei/EA diritti/S umani/A da/E parte/S della/EA polizia/S ,/FF ma/CC su/E scala/S minore/A rispetto/S agli/EA anni/S precedenti/A ./FS
Una/RI rivolta/S carceraria/A nella/EA Prigione/S centrale/S di/E Luanda/S ha/VA determinato/V morti/S e/CC feriti/S ,/FF sebbene/CS le/RD cifre/S a/E tal/DI riguardo/S siano/V state/VA contestate/V ./FS

Tabella (5): Esempio del corpus italiano taggato con Felice-POS-Tagger

Come si può osservare il tagset italiano è tanto semplice rispetto a quello utilizzato da Amira nel corpus arabo. Questa semplicità non ha richiesto tanta normalizzazione, tranne il caso di SP (i nomi con la prima lettera maiuscola) che viene trasformato in S.

183 Dell'Orletta F., "Ensemble system for Part-od-Speech tagging". In *Proceedings of EVALITA 2009 – Evaluation of NLP and Speech Tools for Italian.* Reggio Emilia, Italy.

3.6.2.4 Allineamento

Nell'ambito del TAL per il processo di allineamento si intende rendere due testi corrispondenti allineati l'uno di fronte all'altro.
Questa fase si configura come un processo essenziale lavorando sui corpora paralleli. L'allineamento viene effettuato normalmente da appositi programmi che si servono di metodi statistici e linguistici per mettere in corrispondenza due unità di testo l'una è traduzione dell'altra. Nel caso dei metodi statistici si utilizzano i calcoli probabilistici della lunghezza delle unità (frasi, parole, caratteri) dei due testi paralleli per stabilire un'adeguata equivalenza tra loro. Inoltre, il metodo statistico si può arricchire con repertori lessicali derivati da dizionari o corrispondenze traduttive prestabilite. In alcuni casi l'utilizzo del metodo ibrido appare più conveniente soprattutto quando si tratta di lingue che hanno sistemi di scrittura significativamente diversi tra loro, come per es. le lingue del nostro corpus.

Per allineare i nostri testi, abbiamo utilizzato *LogiTerm* che fa parte di Terminotix[184]. Questo programma segmenta e allinea automaticamente due testi creando il risultato in formati diversi (HTML, XML, TMX). L'accuratezza dell'allineamento nel nostro caso è all'incirca del 95%, quindi è servito un intervento manuale per correggere alcuni errori dovuti in generale alle caratteristiche linguistiche delle due lingue. La maggior parte degli errori individuati durante l'allineamento riguarda la lunghezza della frase araba. Come si può osservare dal numero totale delle frasi nella Tabella (1), la lingua araba tende a congiungere le frasi, quindi non è raro di trovare un livello di allineamento del tipo 2 a 1. Dopo la verifica manuale dei risultati di questa fase, i testi allineati sono stati salvati in due formati XML e TMX.

184 http://www.terminotix.com/index.asp?lang=en

<prop type="ltattr-match">1-1</prop>
<prop type="ltattr-id">17</prop>
<tuv xml:lang="it">
<seg>Ogni persona ha diritto al godimento dei diritti e delle libertà riconosciuti e garantiti nella presente Carta senza alcuna distinzione, in particolare senza distinzione di razza, sesso, etnia, colore, lingua, religione, opinione politica o qualsiasi altra opinione, di origine nazionale o sociale, di fortuna, di nascita o di qualsiasi altra situazione.</seg>
</tuv>
<tuv xml:lang="ar">
<seg>ytmtE kl $xS bAlHqwq wAlHryAt AlmEtrf bhA wAlmkfwlp fY h*A AlmyvAq dwn tmyyz xASp <*A kAn qA}mA ElY AlEnSr >w AlErq >w Allwn >w Aljns >w Allgp >w Aldyn >w Alr>Y AlsyAsY >w >Y r>y |xr‹ >w .Almn$> AlwTny >w AlAjtmAEy >w Alvrwp >w Almwld >w >Y wDE |xr
</seg>
</tuv>
</tu>
<tu>

Tabella (6). Esempio del corpus allineato in TMX

<seg match="1-1" id="17">
<src>Ogni persona ha diritto al godimento dei diritti e delle libertà riconosciuti e garantiti nella presente Carta senza alcuna distinzione, in particolare senza distinzione di razza, sesso, etnia, colore, lingua, religione, opinione politica o qualsiasi altra opinione, di origine nazionale o sociale, di fortuna, di nascita o di qualsiasi altra situazione.</src>
<tgt>ytmtE kl $xS bAlHqwq wAlHryAt AlmEtrf bhA wAlmkfwlp fY h*A AlmyvAq dwn tmyyz xASp <*A kAn qA}mA ElY AlEnSr >w AlErq >w Allwn >w Aljns >w Allgp >w Aldyn >w Alr>Y AlsyAsY >w >Y r>y |xr‹ >w Almn$> AlwTny >w AlAjtmAEy >w Alvrwp >w Almwld >w >Y wDE |xr.
</tgt>
</seg>

Tabella (7). Esempio del corpus allineato in XML

Capitolo IV: Estrazione dei termini monolingui

4.1. Estrazione automatica di termini da corpora

Lo sviluppo crescente nel campo della tecnologia informatica ha consentito una produzione testuale esponenziale. Parallelamente a questa quantità di informazioni che cresce in un modo resistente aumenta "la necessità di strumenti che consentano di operare su di esse in modo automatico, efficace, economico"[185]. Da qui nasce il termine *Text Analysis* (TA) che, diversamente dal termine *analisi testuale* che si basa effettivamente sui mezzi non automatici per effettuare l'analisi del testo, concerne l'analisi automatica di un testo soprattutto di una certa dimensione[186]. L'idea di ancorarsi ai linguaggi formali per effettuare analisi testuali è stata suggerita appunto dall'osservare che nelle lingue umane i fenomeni linguistici possono avere una certa essenza probabilistica individuabile tramite le regole statistiche.

Una *regola linguistica* può essere vista come la descrizione di una pratica linguistica. La descrizione di tali pratiche può avere tuttavia implicazioni più o meno forti. La posizione più debole sostiene che la regola linguistica rappresenti una semplice regolarità, una tendenza che preferisce determinate soluzioni in una lingua, rispetto ad altre meno frequenti. In questo senso la regola è rappresentabile mediante strumenti statistici o probabilistici (tipici nella descrizione delle scienze della vita.)[187]

Secondo Chiari, la nozione di linguistica quantitativa o linguistica statistica risale a tempi remoti: già i greci e romani si accorsero della significativa differenza nella distribuzione della frequenza delle parole nella lingua, arrivando a distinguere tra le parole di alto uso e quelle di bassa frequenza fino agli *hapax*, cioè le parole con una sola occorrenza nei corpora[188].

Tra i vari approcci della linguistica quantitativa appare quello *statistico-descrittivo* che mira " all'estrazione di regolarità statistiche da grandi quantità

185 Zampolli A. (1997). Introduzione. In P. Ridolfi e R. Piraino (a cura di), *Trattamento automatico della lingua nella Società dell'informazione*, Roma, 13/14 gennaio (Aula Magna) Ministero P.T., Atti della Conferenza. *La Comunicazione*, XLVI, numero unico (1,2,3,4). p.1.
186 Bolasco, S., et al., "Estrazione automatica d'informazione dai testi". In *Mondo digitale*, n.1 marzo 2004
187 Chiari, I., *Introduzione alla linguistica computazionale*, Roma, Editore Laterza, 2007, pp 8:9
188 Ibidem

di raccolte testuali"[189]. La fondazione di questo approccio viene attribuita a Zipf, con la sua famosa legge di Zipf, che ha cercato di studiare le lingue umane mediante l'applicazione di principi statistici per arrivare ad una metodologia simile a quella adottata nelle scienze esatte. Secondo Zipf i fenomeni linguistici nei testi come la lunghezza o la brevità delle parole in un testo hanno qualche rapporto con la loro frequenza di occorrenza nel corpus, in un tentativo di guardare alle regolarità statistiche nei testi non come una specie di caso, bensì come "delle caratteristiche di finitezza psico-biologica dell'essere umano" [190].

Sviluppare metodi sofisticati per estrarre termini da corpora rappresenta un'importanza assoluta in molte applicazioni del TAL, soprattutto in compiti come *l'information retrieval* (*reperimento delle informazioni*) e la costruzione delle ontologie, in quanto "both Ontology Learning and Semantic Web technologies often rely on domain knowledge automatically extracted from corpus through the use of tools able to recognize important concepts, and relations among them, in form of terms and terms relations"[191]. L'estrazione dei termini esercita un ruolo principale, per di più, nello sviluppare dei programmi efficaci e accurati nel campo della traduzione automatica, in quanto in questo caso più ampie sono le risorse lessicali da cui dipende un programma di traduzione automatica più accurata è la sua performance. Un ulteriore campo legato all'estrazione terminologica è *l'Automatic Keyword Extraction*, (l'estrazione automatica delle parole chiave dai documenti), definito come "the task to identify a small set of words, key phrases, keywords, or key segments from a document that can describe the meaning of the document".[192] Tuttavia, in quest'ultimo compito vengono prese in considerazione solo le parole chiave e non tutti i termini[193].

---

189 Ivi, p.36

190 Ivi, p.37

191 Pazienza, M.T., et al.,: "Terminology extraction: an analysis of linguistic and statistical approaches". In *Knowledge Mining*, Springer Verlag, 2005, p.255

192 Zhang, C. et al., "Automatic keyword extraction from documents using conditional random fields". In *Journal of computational and information systems* 4:3, 2008, p.1169

193 Cfr. Foo. J.,*Computational Terminology: Exploring Bilingual and Monolingual Term Extraction*. Licenciate Thesis. Linköping University, Department of Computer and Information Science, NLPLAB - Natural Language Processing Laboratory, 2012

4.2. Metodi di estrazione

I metodi adottati nel campo dell'estrazione dei termini consistono in tre tipi: approcci linguistici, approcci statistici e altri ibridi. Mentre gli approcci linguistici dipendono dall'annotazione linguistica, o specificamente morfo-sintattica (PoS tagging), per individuare termini candidati attraverso certi pattern sintattici, quelli statistici si avvalgono di diverse misure di associazione lessicale per determinare, in una fase, il grado di coesione e di connessione tra le componenti dei termini composti, e estrarre, in una fase successiva, solo quei termini i cui costituenti presentano una certa rilevanza dal punto di vista statistico. Un sistema ibrido di estrazione si serve, invece, di entrambi i metodi precedenti per garantire una performance ottimale.

4.2.1. Approcci linguistici

L'approccio linguistico adottato per l'estrazione delle collocazioni in generale e delle terminologie in particolare si basa di solito su due moduli: *parsing module* e *term recogniser module*[194]. Mentre nel primo modulo si tenta di assegnare ad ogni parola nel corpus la sua categoria grammaticale nel contesto mediante il PoS tagging, il secondo modulo si utilizza per filtrare, attraverso regole sintattiche o vari linguaggi formali, le parole del corpus selezionandone solo quelle che possono rappresentare termini candidati.

Si tratta, quindi, di approcci specificamente legati al sistema linguistico o al dominio sociolinguistico di interesse, in quanto ogni lingua ha le sue caratteristiche morfo-sintattiche, dette *synaptic compositions*[195], per formare le proprie terminologie e collocazioni. Inoltre, le differenze morfo-sintattiche possono avvenire anche a livello intralinguistico, in base, quindi, al linguaggio settoriale o specialistico in questione, come per es. il linguaggio giuridico che presenta soluzioni morfo-sintattiche ben differenti da quelle offerte da altre lingue speciali.

194 Pazienza, M. T., et al., "Terminology extraction", op cit, p. 256
195 Ivi, p.257

4.2.2. Metodi statistici

Utilizzare il metodo statistico nell'analisi linguistica presenta più di un vantaggio. Dal momento che si tratta di "un approccio puramente formale che privilegia i segni (significanti) per arrivare al senso (in quanto insieme di significati)", l'approccio statistico si può utilizzare indipendentemente dalla lingua dell'analisi, dall'ampiezza o dimensione della raccolta testuale[196].

Malgrado la comune tendenza a basarsi sui termini composti nella maggior parte dei lavori dell'estrazione terminologica, i termini semplici o monorematici rappresentano a loro volta una rilevanza, perché anche loro possono fungere da potenziali termini nonché possono aiutare a identificare le unità lessicali polirematiche :

A number of existing approaches focused solely on the identification of complex terms. Even though complex terms are more highly regarded in terminology, one cannot deny the fact that simple terms have a role to play. In fact, besides being potential terms, simple terms have the capability of assisting in the identification of complex terms through the head-modifier principle. For an approach to be considered as practical for real-world applications, simple terms cannot be neglected.[197]

Estrarre le unità lessicali composte viene effettuato normalmente nell'ambito della linguistica computazionale in base alla nozione dell'associazione lessicale tra i costituenti del termine. Si parla quindi di una serie di misure statistiche di cui si serve la linguistica computazionale per quantificare la forza di legame tra due o più unità lessicali o in altri termini "interpretare" questo legame di associazione in algoritmi computazionali. Il principio di base di questi metodi statistici parte dall'idea che se due o più unità lessicali compongono una collocazione o una terminologia, è molto probabile che esse co-occorrano di solito nei testi in maniera statisticamente considerevole. Questo fenomeno si può spiegare considerando il fatto che

196 Bolasco, S., "Statistica testuale e text mining: alcuni paradigmi applicativi". In *Quaderni di statistica*, vol.7, 2005, p.20

197 Wong, W., et al., "Determination of Unithood and Termhood for Term Recognition" in *Handbook of Research on Text and Web Mining Technilogies,* Vol. 2, IGI Global, USA, 2009, p. 504.

"le collocazioni sono strutture prefabbricate che i parlanti tendono a usare come blocchi linguistici tendenzialmente unitari e fissi. Per identificare le collocazioni in un corpus basta, dunque, analizzare il modo in cui si distribuiscono le sue parole[198].

Nell'ambito dell'estrazione dei termini si deve prestare attenzione ai seguenti due concetti: *unithood* e *termhood*. Mentre il primo vuol dire il grado di forza o di stabilità di combinazioni sintagmatiche o collocazionali, il secondo si riferisce al grado di rilevanza di un'unità linguistica nei riguardi di un concetto nell'ambito di un dominio specifico[199]. Ciò vuol dire che mentre il termine *unithood* è legato solo alle unità lessicali composte, come le unità terminologiche poliremathiche, le espressioni idiomatiche e le collocazioni, *termhood* riguarda anche le unità monorematiche.

I quanto segue esponiamo in modo breve le principali misure statistiche adoperate per l'estrazione terminologica.

4.2.2.1. Frequenza

I primi tentativi statistici che si occupavano di estrarre collocazioni da corpora testuali utilizzavano la frequenza, cioè il conteggio di distribuzione delle occorrenze di certi unigrammi, bigrammi o trigrammi di parole per individuare le sequenze di parole che probabilmente costituiscono una collocazione. Choueka[200] è stato fra i primi ad utilizzare questo metodo nel suo corpus come parametro per individuare le sequenze di parole adiacenti da un alto valore di co-occorrenza. Si tratta di un approccio semplice che non richiede componenti morfologici o sintattici. Un modo per migliorare i risultati dell'approccio di Choueka è di associare alla frequenza delle parole adiacenti anche informazioni morfo-sintattiche, fornite dal PoS tagging[201],

198 Lenci A., et al., 2012, op cit, p.198

199 Kageura, K., Umino, B., "Methods of automatic term recognition". In *Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication*, 3, 1996

200 Choueka, Y., "Looking for needles in a haystack or locating interesting collocational expressions in large textual databases". In *Proceedings of the RIAO Conference on User-Oriented Content-Based Text and Image Handling*, Cambridg, MA, March 21-24, 1988

201 Justeson J.S., Katz. S.M., "Technical terminology: some linguistic properties and an algorithm for identification in text". In: *Natural Language Engineering,* 1:9-27, 1994

con cui si può creare una specie di filtro linguistico per eliminare, cioè, quelle sequenze di parole che non rientrano nella fascia delle collocazioni interessate come le preposizioni, gli articoli, ecc. nonché focalizzare la ricerca di termini solo sulle frasi sintagmatiche, ecc..

Tuttavia, il metodo di frequenza deve prendere in considerazione anche le diverse variazioni dei termini a livello del corpus testuale perché può succedere che il valore effettivo di occorrenza di un termine non corrisponda al calcolo statistico ottenuto come parametro di frequenza, dal momento che un solo termine può avere varie forme di occorrenza. In un esperimento effettuato su un corpus di testi chimici di 30,000 parole, Le An Ha[202] ha rilevato che un termine come *hydrochloric acid* presenta una frequenza di 1 occorrenza, mentre dopo l'aggiunta delle altre sue variazioni, comprese le espressioni anaforiche, la frequenza del termine sale a 11 occorrenze. Questo vuol dire che la frequenza non si configura come un metodo efficace nel campo dell'estrazione dei termini se non tiene conto delle diverse variazioni dei termini, che risulta, infine, un compito abbastanza complicato nell'ambito del TAL, perché richiede, oltre alle informazioni lessicali e sintattiche, altre conoscenze semantiche nonché pragmatiche[203].

Inoltre, il metodo di frequenza non si considera tanto conveniente nell'identificare sequenze di unità lessicali da corpora specialistici in cui un termine può non ricevere alto valore di occorrenza, e quindi non sarà estratto con l'approccio di frequenza.

4.2.2.2. Media e Varianza

Il metodo della frequenza per estrarre le collocazioni funziona abbastanza bene per quelle sequenze lessicali che presentano una certa *atomicità sintattica*[204], quelle espressioni, cioè, che in contesto mantengono quasi sempre il loro ordine interno nonché la distanza intercorrente tra le

202 Le An Ha (2007) *Advances in Automatic Terminology Processing: Methodology and application in focus*. PhD Thesis, University of Wolverhampton, UK , p.48

203Cfr. Mitkov, R., et al., "A new, fully automatic version of Mitkov's knowledge poor pronoun resolution method". In *Proceedings of CICLing*, Mexico City, Mexico, 2002

204 Grossman, M, Rainer, F., *La formazione delle parole in italiano*, op cit, p. 33

componenti. Tuttavia non tutte le collocazioni o le terminologie presentano questa caratteristica di atomicità sintattica, ma ce ne sono tante che si possono distribuire liberamente a livello di frase, il che significa che il metodo della frequenza non sarà di grande aiuto in questi casi. Pertanto si procede a calcolare approssimativamente la distanza che intercorre tra gli elementi dei termini tramite il calcolo della media e della varianza di queste distanze. Mentre la prima misura il valore medio delle parole che possono intercorrere tra le parole candidate, la seconda stima quanto le distanze si allontanino dal valore medio. Questo nuovo metodo consente di acquisire una varietà molto ampia di collocazioni. Smadja[205] ha sviluppato un sistema di estrazione (Xtract) molto sofisticato e accurato rispetto agli approcci che ci erano in quel tempo. Il nuovo sistema si basa sulla frequenza di co-occorrenza e sulla distanza tra base e collocato. L'idea è che le parole che presentano qualche relazione di associazione o correlazione lessicale risultano regolarmente vicine a livello dell'asse sintagmatico. Inizialmente si calcola il valore medio di distanza fra le parole in questa maniera: $\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$

dove d_i è ogni distanza, cioè il numero delle parole che intercorrono tra la base e il collocato, mentre n è il numero delle co-occorrenze delle due parole.

La varianza, invece, si calcola come segue: $s^2 = \frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n-1}$

dove n è il numero delle volte che le due parole co-occorrono nel corpus, d_i è ogni distanza dell'occorrenza, mentre $\bar{d}$ è la media delle distanze. Se d_i è sempre uguale in tutti i casi, questo significa che il valore della varianza è zero. Se invece le d_i risultano casualmente distribuite (in caso per esempio di parole che si presentano insieme per caso senza che ci sia una relazione di collocazione), la varianza ha un valore più alto.

205 Smadja, F., "Retrieving collocations from text: Xtract", in *Computational Linguistics*, Cambridge, MIT Press, 19(1), 1993

Per stimare la variabilità delle distanze tra due parole si usa la deviazione standard che è la radice quadrata della varianza $\sqrt{s^2}$.

La media e la deviazione descrivono la distribuzione delle distanze tra due parole in un corpus, quindi si può utilizzare questa informazione nell'identificazione di collocazioni guardando alle coppie di parole con un valore basso di deviazione standard. Un valore basso significa che le parole co-occorrono di solito alla stessa distanza. Una deviazione a valore zero significa che le parole si presentano nei testi sempre con la stessa distanza. Allo stesso modo le parole con un alto valore di deviazione non possono diventare buone candidate di collocazioni.

Variance-based collocation discovery is the appropriate method if we want to find this type of word combination, combinations of words that are in a looser relationship than fixed phrases and that are variable with respect to intervening material and relative position[206].

4.2.2.3. Hypothesis testing

L'alta frequenza e la bassa varianza di una coppia di parole potrebbero anche non essere un significativo rilevatore della loro interdipendenza lessicale, quando succede per esempio che le due parole in questione rientrino nella fascia del lessico tanto frequente nel corpus. Pertanto per decidere se due parole possono rappresentare una collocazione o una terminologia dobbiamo stabilire se la loro co-occorrenza è un fatto accidentale o meno. Il test di verifica di ipotesi (hypothesis testing) risulta di grande importanza in questi casi. Primariamente si formula un'ipotesi nulla H_o che non ci sia un'associazione tra le parole selezionate, il che vuol dire che la loro compresenza nel corpus è solo accidentale. Si calcola poi la probabilità della H_o, e in base al valore di questa probabilità si può accettare o rifiutare l'ipotesi nulla. Esponiamo di seguito i test più importanti per quantificare questa probabilità.

206 Manning, C., Schütze H. (a cura di) *Foundations of Statistical Natural Language Processing*, Cambridge, Massachusetts, MIT Press, 1999, p:152

4.2.2.3.1. T-test

Il t-test considera la media e la varianza di un campione, dove l'ipotesi nulla è che il campione sia caratterizzato da un determinato valore di media. Il test osserva la differenza tra le medie osservate e quelle previste, classificando i valori in base alla varianza dei dati per decidere quanto sia probabile raggiungere il valore del campione esaminato. La formula del t-test è la seguente:

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

dove $\bar{x}$ è la media del campione, s^2 è la varianza del campione, N è la dimensione del corpus, e μ è la media della distribuzione.

Rispetto al valore di confidenza del t-test, se t ha un valore abbastanza alto si può rifiutare l'ipotesi nulla e considerare valido quindi il legame di associazione tra le parole selezionate, mentre nel caso contrario si accetta l'ipotesi nulla consistente nel ritenere accidentale la co-occorrenza tra le parole.

4.2.2.3.2. Chi -Square test

Un altro test utilizzato per l'associazione tra le parole è il χ^2 test (chi-square test). Nella sua applicazione il χ^2 test si applica a tabelle 2 × 2 di frequenza. Il test confronta le frequenze osservate con quelle attese per l'indipendenza (cioè la mancanza di associazione tra le componenti di una collocazione). In base al valore di differenza tra le frequenze osservate e quelle attese si procede ad accettare o a respingere l'ipotesi nulla di indipendenza. La formula del test χ^2 è la seguente:

$$x^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

dove i per le righe e j per le colonne della tabella, O_{ij} per i valori osservati di i e j, e E_{ij} per i valori attesi di i e j.

Mentre i valori osservati sono le occorrenze registrate delle parole nel corpus, i valori attesi vengono calcolati dalla probabilità marginale, cioè dal totale delle righe e delle colonne. Per esempio il valore atteso del bigramma w^1 w^2 è il prodotto delle probabilità marginali di w^1 e w^2, diviso per il numero dei bigrammi nel corpus.

Un ulteriore uso del χ^2 è nell'estrazione delle unità di traduzione corrispondenti da un corpus parallelo [207].

4.2.2.3.3. Log likelihood ratio (LLR)

LLR[208]è un test statistico utilizzato ampiamente nel campo di test di verifica di ipotesi. Nei lavori dell'estrazione terminologica il LLR appare appropriato a estrarre termini a bassa frequenza in corpora. Nel caso delle associazioni lessicali si può pensare al LLR come la relazione di verosimiglianza delle statistiche fornite dal corpus d'acquisizione per accettare o respingere l'ipotesi dell'interdipendenza tra due o più parole. Si formulano così due ipotesi[209]:

- Hypothesis 1 (indipendenza). $p(w^2|w^1) = p = p(w^2|\neg w^1)$
- Hypothesis 2 (dipendenza). $p(w^2|w^1) = p_1 \neq p_2 = p(w^2|\neg w^1)$

Si utilizza poi il metodo della massima verosimiglianza per p, p_1 e p_2 :

$$p=\frac{c_2}{N}, p_1=\frac{c_{12}}{c_1}, p_2=\frac{c_2-c_{12}}{N-c_1}$$

dove c_1 è il numero delle occorrenze di w^1; c_2 è il numero delle occorrenze di w^2; c_{12} è il numero delle occorrenze di w^1 w^2 insieme; N è il numero dei termini nel corpus.

$$L(H_1) = b(c_{12}; c_1, p)b(c_2 - c_{12}; N - c_1, p)$$

207 Ivi, p:160

208 Dunning, T., "Accurate Methods for the Statistics of Surprise and Coincidence". In *Computational Linguistics*, 19(1), 1993.

209 Manning, C., Schütze H., *Foundations of Statistical Natural Language Processing*, op cit, p. 172

$L(H_2) = b(c_{12}; c_1, p_1)\, b(c_2 - c_{12}; N - c_1, p_2)$

Il log likelihood ratio λ si calcola poi in questa maniera:

$$\log \lambda = \log \frac{L(H_1)}{L(H_2)}$$

$$= \log \lambda = \log \frac{b(c_{12}; c_1, p)\, b(c_2 - c_{12}; N - c_1, p)}{b(c_{12}; c_1, p_1)\, b(c_2 - c_{12}; N - c_1, p_2)}$$

$$= \log L(c_{12}; c_1, p) + \log L(c_2 - c_{12}; N - c_1, p)$$
$$- \log L(c_{12}; c_1, p_1) - \log L(c_2 - c_{12}; N - c_1, p_2)$$

dove $L(p; n, k) = k \log(p) + (n - k) \log(1-p)$

4.2.2.4 Mutua informazione (MI)

La *mutua informazione* è fra le misure statistiche standard per quantificare il legame di associazione tra le unità lessicali in un testo. Date due parole tipo w^1 e w^2 la MI confronta la probabilità di osservare il bigramma (w^1 w^2) con la probabilità di osservare le componenti dello stesso bigramma l'una indipendente dall'altra. MI ha quindi la formula seguente:

$$MI(w^1, w^2) = \log_2 \frac{P(w^1, w^2)}{P(w^1) P(w^2)}$$

Questo vuol dire che MI (w^1, w^2) misura quanto sia indicativa la presenza di w^1 nel testo per aspettare che w^2 appaia subito dopo e viceversa. Tuttavia, la MI sembra un metodo conveniente per misurare il grado di indipendenza tra le parole piuttosto che il loro grado di associazione o dipendenza, in quanto uno dei problemi di questo metodo statistico è che tende ad essere molto alto per le parole che sono molto rare nel corpus (hapax), dal momento che dipende dalla frequenza delle singole parole.

4.2.2.5 TF-IDF

Un'altra misura statistica adottata per l'estrazione di termini da corpora è TF-IDF che sta per "Term Frequency, Inverse Document Frequency". Il test misura la specificità o la rilevanza terminologica di una collocazione lessicale in base alla sua frequenza in una collezione diversa di documenti. Il valore di TF-IDF aumenta in modo proporzionale al numero dell'occorrenza del termine in un documento, ma nello stesso tempo cresce inversamente in proporzione all'occorrenza dello stesso termine nella collezione di altri documenti. Due sono i componenti di questo metodo : TF (*Term Frequency*) quantifica la frequenza di un termine in un documento nella maniera seguente:

$$TF(t, d_i) = \frac{n_{t,i}}{\sum_{k=1}^{|T|} n_{k,i}}$$ [210],

dove $TF(t, d_i)$ è la frequenza del termine t in un documento d_i ; $n_{t,i}$ è il numero di occorrenza del termine t in un documento d_i; $n_{,k,i}$ è il numero di occorrenza di tutti i termini in un documento d_i.

L'altro fattore della funzione (IDF) misura l'importanza o la significatività di un termine in una collezione di corpora:

$$IDF_t = \log \frac{M}{m_t + 0.01}$$

dove M = il numero totale dei documenti del corpus e m_t è il numero dei documenti in cui appare il termine candidato.

E per misurare, infine, il peso (w) del termine t nel documento d_i ci si serve della formula: $w(t, d_i) = TF(t, d_i) \times IDF_t$ [211]

210 Chakraborty, R., "Domain Keyword Extraction Technique: a new weighting method nased on frequency analysis". In *ACER,* 2013, p.111
211 Ibidem

4.2.2.6 C-NC value

C-NC value[212] verifica il valore di *termhood* di termini candidati, servendosi delle loro caratteristiche statistiche le quali sono: il numero di occorrenza all'interno del corpus; il termine annidato (*term nested*) che significa la frequenza del termine candidato come partecipe di altri termini più lunghi; il numero di quei termini più lunghi; e la lunghezza del termine candidato.

$$\text{C-Value(a)} = \begin{cases} \log_2 (|a|) \cdot f(a) & \text{se } a \text{ non è annidato,} \\ \log_2 (|a|) \cdot \left(f(a) - \frac{1}{p(T_a)} \sum_{b \in T_a} f(b)\right) & \text{altrimenti} \end{cases}$$

dove *a* è il termine composto, $|a|$ è la lunghezza in parole di *a*, *f(a)* è la frequenza di *a* all'interno del corpus, T_a è la serie dei termini contenenti *a*, $p(T_a)$ è il numero dei termini in T_a. Come si vede, se il termine candidato non risulta annidato, cioè non fa parte di altri termini più lunghi, il suo valore di *termhood* dipenderà solo dalla sua frequenza di occorrenza nel corpus e dalla sua lunghezza. Se invece il termine candidato appare annidato, la sua *termhood* prenderà in considerazione la sua frequenza come parte di altri termini nonché il numero di questi termini in cui appare.

La funzione NC-value combina il valore di C-value di un termine candidato con le sue informazioni contestuali che si basano sulle parole che appaiono vicine al termine nel contesto. A seconda del numero dei termini con cui appare, una parola si può considerare parola di contesto. Il peso di una parola di contesto si calcola nel modo seguente: $weight(w) = \frac{a(w)}{n}$

dove *w* è la parola di contesto; *a(w)* è il numero dei termini in cui *w* appare; *n* è il numero totale dei termini candidati. Quindi la funzione N-value del termine *a* è definita come segue: $N\,value\,(a) = \sum_{w \in C_a} f_a(w) * weitht(w)$

dove $f_a(w)$ è la frequenza di *w* come parola di contesto con il termine *a* e C_a è la collezione delle parole di contesto di *a*. Combinato con C-value, il valore N-value dà luogo alla funzione NC-value.

212 Frantzi K., Ananiadou S.,"The C–value / NC Value domain independent method for multi–word term extraction". In *Journal of Natural Language Processing*, 6(3), 1999

4.3. Stato dell'arte dell'estrazione di termini da corpora italiani e arabi

In questa parte passiamo per rassegna gli importanti tentativi dell'estrazione terminologica da corpora monolingue sia italiani che arabi.

4.3.1. Estrazione di termini da corpora italiani

In quanto segue esponiamo i lavori più importanti dell'estrazione di termini da corpora italiani.

1- Partendo dall'idea che l'individuazione dei termini dovrebbe basarsi non solo sulle informazioni distribuzionali delle parole nel testo, bensì su altri indizi extra-contestuali derivati da altri domini, nel 2001 Basili et al[213] hanno presentato un approccio contrastivo per estrarre termini da corpora di dominio. La base del loro approccio dipende dal concetto principale secondo il quale "better models should be derived over different samples spaces rather than in the refinement in the probabilistic measures in the target domain"[214]. Questo vuol dire che riconoscere la specificità dei termini e, quindi, estrarli dal loro contesto non deve essere affidato solo al principio di frequenza di tali termini nel documento, bensì al confronto di questa loro frequenza con altri domini specifici.

L'approccio consiste di due componenti:

- analisi sintattica superficiale (shallow parsing): l'analisi linguistica, detta anche *the symbolic feeder*, dipende dal parser Chaos[215] che fornisce un insieme di moduli per il trattamento automatico del linguaggio, quali: a) tokenizzatore; b) *yellow page look-up module*, che abbina i nomi di entità esistenti in cataloghi; c) analizzatore morfologico che assegna, con possibili ambiguità, ad ogni parola nel testo la sua categoria sintattica e la sua interpretazione morfologica; d) *named entities matcher* che riconosce,

213 Basili, R., et al., "A contrastive approach to term extraction". In *Proceedings of the 4th Terminology and Artificial Intelligence Conference (TIA)*, France, 2001

214 Ivi, p.119

215 Basili, R., et al., "Customizable Modular Lexicalized Parsing", in *Proceedings of the 6th International Workshop on Parsing Technology*, IWPT, 2000

secondo certe grammatiche, i nomi di entità complessi; e) *PoS tagger* a regole; f) modulo PoS di disambiguazione, che risolve le eventuali ambiguità morfologiche; g) analizzatore sintattico per costruire i *chunk* del testo analizzato.

Le strutture sintattiche sempliche, *chunk*, costituiscono, poi, il nucleo centrale per acquisire, in base alla nozione di "testa" nei costituenti del complesso nominale o *partial phrase*, unità terminologiche candidate da classificare in candidati semplici e candidati complessi.

- filtro statistico: questa parte nel sistema mira a escludere quei termini candidati che non rappresentano significativi termini di dominio - in questo caso si tratta di un dominio giuridico - fornendo, mediante misure comparative statistiche, liste classificate per ogni dominio. L'input di questa fase sono due liste di termini candidati: una per le unità monorematiche e l'altra per quelle polirematiche.

Per classificare i termini monorematici, viene utilizzato qui il metodo Inverse Word Frequencies (IWF)[216], definto come segue:

$IWF(t)=\log(\frac{N}{F_t})$ [217], dove N è la dimensione del corpus, calcolata come la somma delle frequenze di tutti i termini candidati in tutti i domini; F_t è la frequenza accumulativa del termine *t* in tutti i domini j, cioè $F_t=\sum_j f_t^j$. Infine il peso contrastivo del termine semplice viene calcolato secondo la formula seguente: $w_t^i=\log(F_t^i)*IWF(t)$ [218].

La classifica dei termini complessi, dipende, invece, da due fattori: la suddetta funzione del peso contrastivo per le unità monorematiche, che viene applicata qui alle teste dei termini polirematici; e la frequenza di quei termini complessi nel dominio in questione. Quindi la funzione di selezione contrastiva via testa (*Contrastive Selection via Heads*) dei termini complessi viene definita nel

216 Basili, R., et al., "A text classifier based on linguistic processing". In *Proceedings of IJCAI 1999. Machine Learning for Information Filtering*

217 La formula di IWF si differenzia da quella IDF già trattata, in quanto la prima conta solo le parole mentre la seconda conta i documenti. Cfr. Foo, J., Computational Terminology: op cit.

218 Basili, R., et al., A contrastive approach to term extraction, op cit, p.125

modo seguente: $cw_{ct}^{i} = w_{h(ct)}^{i} \cdot f_{ct}^{i}$ [219], dove f_{ct}^{i} è la frequenza dei termini composti nel corpus, e $w_{h(ct)}^{i}$ è la funzione contrastiva del peso per le loro relative teste.

2- Text-to-Knowledge (T2K): si tratta di un programma sviluppato congiuntamente dall'Istituto di Linguistica Computazionale (CNR) e dal Dipartimento di Linguistica dell'Università di Pisa, rivolto all'estrazione di diversi tipi di informazione semantico-lessicale da corpora testuali.

Attraverso l'uso combinato di tecniche statistiche e di strumenti avanzati per il TAL, T2K è in grado di analizzare il contenuto linguistico dei documenti, individuare i termini potenzialmente più significativi, ricostruire una "mappa" multidimensionale dei concetti espressi da questi termini, sviluppare un'ontologia del dominio di interesse[220].

La piattaforma T2K si articola in due livelli:

- glossario terminologico: in questa fase si utilizzano le annotazioni morfo-sintattiche e sintattiche per estrarre automaticamente termini mono e polirematici. Ai termini estratti vengono applicati i diversi metodi statistici per selezionarne i più rilevanti e significativi dal punto di vista di *unithood*. Nel caso dei termini monorematici la selezione si basa sulla frequenza dei termini lemmatizzati all'interno del documento d'acquisizione. I termini complessi si identificano, invece, sulla base dei testi segmentati sintatticamente in elementi sintattici basilari, *chunk*. Inizialmente si scelgono i chunk candidati a rappresentare veri termini, come per es. la sequenza di chunk nominale (N_C) seguito da un chunk aggettivale (ADJ_C) (es. *organizzazione internazionale*). Alle sequenze di chunk estratte viene applicato il metodo "log-likelihood" per determinare la significatività statistica della loro associazione.

- strutturazione concettuale dei termini: l'organizzazione concettuale consiste

219 Ibidem

220 Dell'Orletta, F., et al., "Dal testo alla conoscenza e ritorno: estrazione terminologica e annotazione semantica di basi documentali di dominio". In *AIDA Informazioni, Atti del Convegno Nazionale Ass.I.Term "I-TerAnDo"*, Università della Calabria, 5-7-giugno 2008, Roma : AIDA, n. 1-2/2008.

nell'identificazione delle relazioni semantiche di iponimia, iperonimia e di affinità semantica tra i termini del glossario estratti nella fase precedente. In questo senso la relazione di inclusione lessicale è la base per identificare relazioni tassonomiche tra i termini composti: due termini polirematici con la stessa testa lessicale vengono considerati iponimi della testa condivisa. Per l'identificazione di affinità semantica si procede, invece, all'assunzione secondo la quale "la semantica di una parola si correla alle sue proprietà distribuzionali nel testo, ovvero due parole sono semanticamente simili se sono reciprocamente sostituibili in un numero significativo di contesti sintattici"[221].

3- In Bonin et al[222] si è cercato di estrarre termini complessi di dominio attraverso un approccio contrastivo che misura la *termhood* di un termine tecnico confrontandone la distribuzione con quella in un corpus generale. Il sistema di estrazione consiste di due fasi: a) estrazione di termini candidati; b) classificare i termini selezionati tramite il confronto con altri corpora. L'estrazione dei termini candidati dipende dal filtro linguistico basato sul PoS tagging del corpus e dal filtro statistico, C-NC value, per selezionare i termini rilevanti. In una seconda fase i termini candidati saranno soggetti ad un ulteriore filtro attraverso il confronto con altri corpora di altri domini. Rispetto agli approcci precedenti, la novità di questo metodo consiste nel fatto che il filtro contrastivo si applica qui ai termini già selezionati. Questo approccio contrastivo si basa su una funzione di arcotangente nel seguente modo: $w(x) = arctan(K * x)$, dove K è il coefficiente. Quindi se T è la collezione dei termini composti estratti dal corpus di acquisizione i e C è la collezione dei corpora di riferimento oppure corpora contrastivi, il coefficiente K è definito nel modo seguente: $K(t) = \frac{1}{\frac{F_c(t)}{N_c}}$

dove $t \in T$, $K(t)$ è il coefficiente di t, $F_c(t)$ è la somma delle frequenze di t all'interno dei corpora di riferimento, N_c è la somma delle frequenze di tutti gli elementi di t all'interno dei corpora di riferimento. Quindi il peso statistico

221 Ivi, p.198

222 Bonin, F. et al, "A Contrastive Approach to Multi-word Extraction from Domain-specific Corpora". In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, 2010, Malta

del termine t = $w(t)=\arctan\left(\frac{\frac{F_i(t)}{F_c(t)}}{N_c}\right)$

dove $f_i\ (t)$ è la frequenza di t all'interno del corpus di acquisizione. Per il problema della bassa frequenza di t all'interno del corpus di riferimento $F_c(t)$, il sistema adottato calcola il valore di Csmw (Contrastive Selection of multi–word terms) moltiplicando l'oggetto dell'arcotangente per il logaritmo della frequenza di t all'interno del corpus di acquisizione, nel modo seguente:

$$CSmw(t)=\arctan\left(\log\left(F_i(t)\right)*\left(\frac{\frac{F_i(t)}{F_c(t)}}{N_c}\right)\right)$$

4.3.2. Estrazione di termini da corpora arabi

In quanto segue riportiamo i più importanti lavori sull'estrazione dei termini da corpora arabi.

1- Al Khatib e Badarneh[223]hanno presentato un approccio all'estrazione di termini complessi dai testi arabi. Il loro sistema si compone di due fasi: una linguistica dove si estraggono termini composti candidati e un'altra statistica ove si usano i due test Log Likelihood Ratio (LLR) e C-Value per filtrare i termini selezionandone solo quelli più significativi. Invece di dipendere dai pattern tradizionali basati sul PoS tagging, questo programma usa nuovi modelli basati sul concetto della definitezza o meno dei nomi, come indicato nella figura seguente:

(1)	Nome definito	⟹	uno o più nomi definiti	
(2)	Nome indefinito	⟹	uno o più nomi indefiniti	
(3)	Nome indefinito	⟹	uno o più nomi definiti	
(4)	(1) o (2) o (3)	⟹	preposizione	⟹ (1) o (2) o (3)

Tabella (8) pattern sintattici utilizzati per estrarre termini composti

223 Al Khatib, K., Badarneh, A., "Automatic extraction of Arabic multi-word terms". In *Proceedings of the International Multiconference on Computer Science and Information Technology*, 2010.

Per classificare morfo-sintatticamente le parole del corpus il sistema utilizza il tagger di Al-Taani[224]. Dopo il PoS tagging si estraggono sequenze nominali connesse con o senza preposizione, come negli esempi seguenti:

- sequenze di nomi connessi senza preposizione: mnZmp Al>rSAd Aljwyp AlEAlmyp (organizzazione meteorologica mondiale)

- nomi connessi con preposizione: AltHkm En bEd (telecomandato)

Per misurare la *unithood* dei termini viene utilizzato il LLR, mentre per la *termhood* viene adottato il test C-Value.

2- Un altro approccio all'estrazione di termini arabi è stato presentato da Bounhas&Slimani[225]. L'estrazione in questo sistema si basa su diversi livelli:

- identificare i confini dei termini complessi per estrarre poi sequenze terminologiche candidate a diventare termini interessanti. Si crea poi una lista di tutte quelle sequenze insieme alla loro relativa frequenza nel corpus d'acquisizione. Si utilizzano, in una fase successiva, delle regole sintattiche per filtrare le frequenze già selezionate.

Il sistema funziona nel modo seguente: primariamente viene avviato l'analizzatore morfologico (AraMorph) parallelamente al PoS tagger. Per integrare tutti e due gli strumenti si sviluppa una terza funzione detta Morpho-PoS Matcher finalizzata, attraverso l'uso di una tabella di corrispondenza tra le etichette dell'analizzatore morfologico e quelle del PoS tagger, a disambiguare l'annotazione morfologica delle parole. L'output di questa fase è una lista delle parole del corpus, ognuna con le proprie possibili soluzioni morfologiche esposte nella maniera seguente: PoS (T, L, D, G, N, A), dove T sta per testo, L per lemma, D per determinatezza, G per genere, N per numero, A per aggettivo. In una fase successiva vengono identificate le sequenze di parole candidate a rappresentare termini. Ogni sequenza contiene le diverse

224 Al-Taani A.T, Abu-Al-Rub S.: "A rule-based approach for tagging non-vocalized Arabic words". In *The International Arab Journal of Information Technology,* Volume6 (3), 2009

225 Bounhas I., Slimani,Y., "A hybrid approach for Arabic multi-word term extraction". In *International Conference on Language Processing and knowledge Engineering*, 2009

caratteristiche morfo-sintattiche e il numero di occorrenza nel corpus. Tale lista di sequenze rappresenta l'input della fase successiva, cioè il parser sintattico che si serve delle regole sintattiche per individuare dei termini nominali complessi. Quelle regole sintattiche hanno il seguente formato:

Adjective_defined (T_1+T_2, L_1+L_2, "DET", G, N_1, 0) ←

NN(T_1 ,L_1 , "DET", G, N_1 ,_), NN(T_2, L_2, "DET", G, N_2, 1)

Questa regola definisce una sequenza di termini nella maniera seguente:

il primo elemento deve avere il PoS NN e deve avere l'articolo determinativo; e il secondo elemento deve a sua volta essere NN e avere l'articolo determinativo e, in più, deve fungere da modificatore aggettivale; infine entrambe le componenti devono concordare in genere.

In un ultimo passo il sistema utilizza la statistica, in questo caso il test LLR, per risolvere le eventuali ambiguità nelle sequenze terminologiche, quando cioè un elemento in una sequenza presenta più di una soluzione morfo-sintattica.

3- Un nuovo approccio all'estrazione di termini arabi è stato presentato da El Mahdaouy et al[226]. Come molti altri sistemi di estrazione, il loro approccio consiste di due parti: una linguistica e un'altra statistica. Nella prima parte linguistica il corpus viene taggato con il sistema di Amira, poi il corpus taggato si tokenizza per identificare, in base a certi pattern morfo-sintattici, delle sequenze nominali candidate a rappresentare unità terminologiche. La sezione linguistica comprende pure l'individuazione delle diverse variazioni terminologiche per migliorare l'efficacia dell'estrazione. La novità di questo approccio appare nella seconda parte, cioè quella statistica mirata a filtrare i termini candidati già selezionati nella parte precedente. Per integrare *termhood* e *unithood* il sistema utilizza il test NTC- Value che si compone dei due metodi C-NC value e T-score, nella seguente formula:

226 El Mahdaouy, A. et al., "A Study of Association Measures and their Combination for Arabic MWT Extraction". In *Proceedings 10th International Conference on Terminology and Artificial Intelligence*, France, 2013

$$F(a) = \begin{cases} f(a) \text{ if } \min(T\ s(a)) \leq 0 \\ f(a)\ \ln(2 + \min(T\ s(a)) \text{ altrimenti} \end{cases}$$ [227]

dove min(T s(*a*)) corrisponde al valore minimo di T-score di tutti i termini candidati *a*. Se si sostituisce F(*a*) alla formula C-value già discussa, si ottiene la nuova formula di TC-value che si combina, alla stessa maniera della C-value, con il metodo di N value, dando luogo alla nuova forma di NTC- value, come segue:

NTC-value(a) = 0.8·T Cvalue(a)+0.2·N value(a)

Similmente il metodo CN-value viene integrato qui con LLR, dando luogo alla nuova forma NLC-value, come segue:

$$\textit{LC-value(a)} = \begin{cases} \log_2(|a|) \cdot FL(a) \text{ se } a \text{ non è annidato,} \\ \log_2(|a|) \cdot (F\ L(a) - GL(a)) \text{ altrimenti} \end{cases}$$

dove FL(a) = f (a) · ln(2 + min(LLR(a))), che viene usato qui al posto di f(a) nell'equazione del metodo C-value; e GL(a) = $\frac{1}{|T_a|}\sum_{b \in T_a} FL(b)$ [228]

Integrato ancora con N-value, quest'ultimo metodo dà luogo al nuovo metodo NLC- value: *NLC-value(a) = 0.8·LC-value(a)+0.2·N value(a)*

4- Attia et al[229] hanno presentato tre approcci complementari all'estrazione di espressioni composte (MWE) da corpora arabi.

Nel primo approccio, chiamato Crosslingual Correspondence Asymmetries, il metodo si concentra sulle MWE non scomponibili semanticamente.

"There are many signs, or indications, of non-compositionality, two well-known among them are "non-substitutability", when a word in the expression cannot be

227 Ivi, p.49
228 Ibidem
229 Attia, M. et al., "Automatic Extraction of Arabic Multiword Expressions", In: *Proceedings of the Workshop on Multiword Expressions: from Theory to Applications*, Beijing, China. 2010

substituted by a semantically equivalent word, and "single-word paraphrasability", when the expression can be paraphrased or translated by a single word."

L'idea principale di questo primo approccio dipende dalla concezione secondo la quale se una MWE in una lingua di origine viene tradotta in una sola parola nella lingua d'arrivo, questo potrebbe rappresentare buon indicatore che siamo davanti ad una MWE caratterizzata dalla non scomponibilità semantica. Questo vuol dire che il tipo di traduzione many-to- one viene impiegato qui per identificare le MWE. In questo caso i ricercatori utilizzano un corpus arabo di titoli raccolti da wikipedia. Dopo la raccolta dei titoli, si effettua la verifica delle corrispondenze traduttive del tipo many-to- one, cioè dall'arabo alle altre lingue, tramite i link interlinguistici forniti da wikipedia.

Il secondo approccio dipende dalla traduzione delle MWE dall'inglese in arabo. Questo approccio si basa sull'ipotesi che una MWE debba essere tradotta nella lingua d'arrivo sempre come MWE. Per fare questo si estraggono le MWE inglesi dal Princeton WordNet[230] e poi si traducono automaticamente, tramite Google Translate, in arabo. Si convalidano poi i risultati ottenuti.

Il terzo e l'ultimo approccio complementare è Corpus-Based Approach. Qui si utilizzano le misure statistiche di associazione lessicale per individuare delle MWE in arabo, e in questo caso si tratta dei test PMI (Pointwise Mutual Information) e Chi-square. Questo ultimo approccio consiste in quattro fasi: a) si contano le frequenze dei unigrammi, bigrammi e trigrammi del corpus; b) si effettuano i calcoli statistici per i bigrammi e trigrammi che poi vengono classificati, in base al loro valore statistico, in ordine discendente; c) si effettua la lemmatizzazione del corpus per gestire le variazioni nelle MWE; d) infine vengono filtrati i risultati tramite il PoS tagging per selezionare quelle sequenze che possono rappresentare vere MWE.

230 http://wordnet.princeton.edu

4.4. Estrazione dei termini dal corpus della tesi

Per estrarre i termini dal nostro corpus, abbiamo adottato un approccio ibrido che combina le informazioni linguistiche fornite dall'analisi morfo-sintattica del corpus (PoS tagging) con i valori statistici offerti dai test statistici.

Il processo di estrazione passa per le seguenti fasi:

4.4.1. Estrazione di termini candidati

In questa fase si utilizzano le caratteristiche morfo-sintattiche assegnate alle parole del corpus nella fase precedente per creare dei pattern sintattici da utilizzare per estrarre dei termini candidati. L'estrazione dei termini candidati viene effettuata, però, dal corpus annotato originale, cioè non lemmatizzato, per garantire una copertura possibilmente esaustiva delle varianti dei termini che servono nell'ultimo capitolo per l'individuazione delle variazioni terminologiche.

I pattern morfo-sintattici delle terminologie giuridiche in arabo si possono riassumere in quanto segue:

- nome + nome: in questo caso il primo nome appare senza l'articolo determinativo "Al", mentre il secondo nome può essere determinato o indeterminato, fungendo così da *retto* nell'ambito dello stato costrutto, come in: (Amr AEtqAl :mandato di arresto ; Hqwq AlAnsAn :diritti umani);

- nome + aggettivo: in questo caso l'aggettivo si accorda sempre al nome precedente in genere, numero, caso e determinazione, come in: (AxlA' qsry : sgombero forzato; AlAxlA' Alqsry :lo sgombero forzato);

- nome + nome + aggettivo: in questo caso l'aggettivo si accorda al secondo nome che a sua volta può essere determinato o indeterminato, come in (thmp AlxyAnp AlEZmY: accusa di alto tradimento; xdmp mdnyp bdylp : servizio civile alternativo);

- nome + nome + nome: in questo caso il terzo nome appare determinato con l'articolo determinativo, mentre i primi due sono indeterminati, come in ($rTp mkAfHp Al$gb : la polizia anti-sommossa);

- nome + nome + nome + aggettivo: (mrD nqS AlmnAEp Almktsbp: HIV)

- nome + nome + nome + nome: (wqf tnfy* >HkAm Al<EdAm: moratoria sulle esecuzioni)

- nome + preposizione + nome: (AlEnf Dd Almr>p : violenze sulle donne)

Così possiamo formulare in modo riassuntivo i principali pattern morfo-sintattici delle terminologie arabe come segue:

– (Nome + (Nome|Aggettivo) + |(Nome|Aggettivo) +|(Nome|Aggettivo))[231]

– Nome + Prep + Nome

In italiano le terminologie presentano i seguenti pattern:

- nome + Prep + nome: prigionieri di coscienza, governo ad interim;

- nome + aggettivo: tribunale civile, perseguimento giudiziario;

- nome + Prep + nome + aggettivo: legge sulla sicurezza interna;

- nome +aggettivo+ Prep + nome + aggettivo: standard internazionali di equità processuale

I precedenti pattern si possono riassumere nel seguente formato:

Nome+(Prep+(Nome|Aggettivo)+|Nome|Aggettivo)+[232]

Il logaritmo di estrazione dei termini candidati cerca nel corpus taggato, a livello monolingue, per trovare le stringhe più lunghe che soddisfino i pattern linguistici prestabiliti. Il codice parte da un certo numero di ngramma per scendere gradualmente finché non arriva ai bigrammi. I termini candidati individuati vengono classificati poi secondo la lunghezza delle stringhe, in modo che in cima alla lista si mettono le stringhe più lunghe e si scende gradualmente fino ai bigrammi.

Le due tabelle seguenti indicano alcuni esempi dei termini candidati estratti.

231 El Mahdaouy A., et al., "A Study of Association Measures and their Combination for Arabic MWT Extraction". In *Proceedings 10th International Conference on Terminology and Artificial Intelligence*, Paris, France, 2013.

232 Bonin, F., et al. "A contrastive approach to multi word term extraction from domain corpora" op cit.

Al>HkAm AlSAdrp En AlmHkmp Aldstwryp (le sentenze della corte costituzionale)
AlmHAkm AlmtxSSp fy AljrA}m Aljnsyp (tribunali specializzati in reati sessuali)
AlmEAmlp Alsy}p fy AlHjz AlEskry (maltrattamenti nelle strutture militari)
AlHZr AlEAm ElY Al>n$Tp AlsyAsyp (divieto generale sulle attività politiche)
AlAqtrAE Alsry fy AljmEyp AlEAmp (scrutinio segreto dall'Assemblea Generale)
AlmwZfyn AlEmwmyyn fy AlqTAE AlxAS (dipendenti pubblici nel settore privato)
AlAst}nAf AlqDA}y Dd AlqrArAt AlSAdrp (appello contro le decisioni assunte)
Alk$f Almbkr En AljrA}m (individuare precocemente i reati)
AlmbAd} AlwArdp fy (principi menzionati nella convenzione internazionale) AlAtfAqyp Aldwlyp
n$Tp mnZmAt AlmjtmE Almdny< (attività delle organizzazioni della società civile)
tqyyd Hryp tkwyn AljmEyAt (limitare la libertà di associazione)
mmArsp AlElAqAt Aljnsyp Almvlyp (praticare l'omosessualità)
hy}p >rkAn AlqwAt AlmslHp (Stato maggiore delle Forze armate)
mrtkbw jrA}m AlAxtfA' Alqsry (perpetratori di sparizioni forzate)
xdmAt AlrEAyp AlSHyp AlmlA}mp (adeguati servizi sanitari)

Tabella (9): Esempi di termini candidati estratti dal corpus arabo

reclutamento di persone a scopo di addestramento
richiesta di risarcimento per tortura in custodia
presa di ostaggi a scopo di riscatto
rapimento di adolescenti a scopo di reclutamento
tratta di donne a scopo di sfruttamento
diritto delle persone alla libertà di circolazione
campo di addestramento militare
forme di discriminazione contro le donne

bande criminali di trafficanti di droga
crimini di violenza contro le donne
uso crudele di dispositivi di contenzione
maltrattamenti di ostaggi durante la prigionia
equipaggiamento di protezione individuale
ispettore generale della polizia
ratifica del Protocollo opzionale
Polizia investigativa criminale provinciale
Commissione verità e riconciliazione

Tabella (10): Esempi di termini candidati estratti dal corpus italiano

4.4.2. Filtro statistico

Questa fase dell'estrazione mira a filtrare i termini candidati estratti nel passo precedente, mediante i test statistici utili a selezionare termini rilevanti e significativi dal punto di vista sia di *termhood* che di *unithood*.

Per assicurare ad ogni termine una concreta distribuzione statistica ci si serve della forma lemmatizzata sia dei termini candidati che del corpus di acquisizione. L'opzione per i lemmi piuttosto che le forme originali dei termini candidati nel processo di classificazione statistica ci è stata suggerita dal lavoro di Felice et al[233]. Il vantaggio derivante da questo approccio è, secondo gli autori, duplice: "da un lato permette di condurre il processo di estrazione terminologica facendo astrazione da variazioni di natura ortografica, morfologica così come strutturale, dall'altro rende possibile l'acquisizione delle varianti terminologiche associate a ciascun termine acquisito".

La classificazione dei termini candidati con le misure statistiche passa per i seguenti passi:

233 Dell'Orletta, F., et al. "Dal testo alla conoscenza e ritorno: estrazione terminologica e annotazione semantica di basi documentali di dominio", op cit.

- trasformare le componenti dei termini candidati in lemma in entrambe le lingue;

- classificare i termini candidati composti secondo la loro lunghezza, in modo che vengano prima i ngrammi, poi n-1grammi fino ai bigrammi;

- utilizzare le misure statistiche per classificare i termini candidati a seconda del grado di associazione lessicale tra i loro costituenti;

4.4.2.1.Metodi di unithood

4.4.2.1.1. Log likelihood ratio (LLR)

Un modo comune per computare i valori di LLR per N-grammi è di creare una tabella di contingenza la quale, nel caso di bigrammi, ha il seguente formato:

	w^1	$\neg w^1$	
w^2	*n_ii*	*n_xi*	*n_pi*
$\neg w^2$	*n_ix*	*n_xx*	*n_px*
	n_ip	*n_xp*	Totale = *n_pp*

Tabella(11): Tabella di contingenza di bigrammi

La cella *n_ii* indica la frequenza congiunta di w^1 w^2; la cella *n_ix* indica il frequenza dei bigrammi in cui w^1 risulta nella prima posizione mentre come secondo elemento non occorre w^2; la cella *n_xi* indica il frequenza dei bigrammi in cui w^2 appare nella seconda posizione mentre come primo elemento non occorre w^1; *n_xx* è la frequenza dei bigrammi le cui componenti non sono né w^1 né w^2; *n_ip, n_xp, n_pi* e *n_px* rappresentano le frequenze

marginali delle occorrenze/non occorrenze delle w^1 e w^2; e infine *n_pp* è il numero totale dei bigrammi nel corpus di acquisizione.

Nel nostro caso abbiamo calcolato il valore LLR utilizzando la formula di G^2, che prende in considerazione i valori osservati e i valori attesi di una tabella di contingenza in questa maniera:

$$G^2 = \quad 2*\sum_i^j O_{ij}*\log(\frac{O_{ij}}{E_{ij}}) \quad ,$$

dove O_{ij} sono i valori osservati di N-grammi nel corpus, mentre E_{ij} sono i loro relativi valori attesi, che rappresentano in questo caso il modello dell'ipotesi di indipendenza. Mentre i valori osservati sono le frequenze registrate di un N-grammi nel corpus, quelli attesi si calcolano dividendo il prodotto della frequenza marginale delle componenti del N-grammi per il numero totale dei N-grammi nel corpus. Quindi i valori attesi della cella (*n_ii*) nella tabella precedente si calcolano in questo modo: *E (n_ii) = (n_ip * n_pi)/n_pp*

Nel caso dei 3-grammi[234] e 4-grammi la tabella di contingenza avrà il seguente formato:

		w^3	$\neg w^3$	
w^1	w^2	*n_iii*	*n_iix*	$n_{\_iip}$
w^1	$\neg w^2$	*n_ixi*	*n_ixx*	$n_{\_ixp}$
$\neg w^1$	w^2	*n_xii*	*n_xix*	$n_{\_xip}$
$\neg w^1$	$\neg w^2$	*n_xxi*	*n_xxx*	$n_{\_xxp}$
		$n_{\_ppi}$	$n_{\_ppx}$	Totale=$n_{\_ppp}$

Tabella(12): Tabella di contingenza di 3-grammi

234 Mcinnes, B.T., *Extending the Log Likelihood Measure to Improve Collocation Identification*, Master's thesis, University of Minnesota, 2004

			w^4	$\neg w^4$	
w^1	w^2	w^3	n_iiii	n_iiix	$n_{\_iiip}$
w^1	w^2	$\neg w^3$	n_iixi	n_iixx	$n_{\_iixp}$
$\neg w^1$	w^2	w^3	n_xiii	n_xiix	$n_{\_xiip}$
w^1	$\neg w^2$	w^3	n_ixii	n_ixix	$n_{\_ixip}$
w^1	$\neg w^2$	$\neg w^3$	n_ixxi	n_ixxx	$n_{\_ixxp}$
$\neg w^1$	w^2	$\neg w^3$	n_xixi	n_xixx	$n_{\_xixp}$
$\neg w^1$	$\neg w^2$	w^3	n_xxii	n_xxix	$n_{\_xxip}$
$\neg w^1$	$\neg w^2$	$\neg w^3$	n_xxxi	n_xxxx	$n_{\_xxxp}$
			$n_{\_pppi}$	$n_{\_pppx}$	Totale=$n_{\_pppp}$

Tabella(13): Tabella di contingenza 4-grammi

E in questo caso i valori attesi delle cella *n_iii* si calcolano in questo modo:

$$E\ (n\_iii) = \frac{(n_{iip}+n_{ixp})*(n_{iip}+n_{xip})*(n_{ppi})}{n_{ppp}} \ ,$$

dove il numeratore è il prodotto delle frequenze marginali delle w^1, w^2, w^3 nelle loro rispettive posizioni, mentre il denominatore è il numero totale dei trigrammi nel corpus.

In una maniera simile si calcola il valore atteso della cella *n_iiii* nella tabella di 4-grammi:

$$E\ (n\_iiii) = \frac{(n_{iiip}+n_{iixp}+n_{ixip}+n_{ixxp})*(n_{iiip}+n_{iixp}+n_{xiip}+n_{xixp})*(n_{pppi})}{n_{pppp}} \ ,$$

dove il numeratore è il prodotto delle frequenze marginali delle w^1, w^2, w^3, w^4 nelle loro rispettive posizioni, mentre il denominatore è il numero totale dei 4-grammi nel corpus.

La selezione dei termini tramite il metodo di LLR viene applicata, a livello monolingue, a tutta la lista dei termini estratti nella fase precedente. I termini con un valore LLR sopra la soglia prestabilita (nel nostro caso è stimata per 10) creano una lista ordinata a seconda del valore LLR.

Le due tabelle seguenti indicano i primi 5 termini delle due liste dei termini selezionati con LLR.

Termine lemma	Termine originale
diritto umano	diritti umani
nazione unito	Nazioni Unite
corte supremo	Corte Suprema
pena detentivo	pena detentiva
sparizione forzato	sparizione forzata

Tabella(14). I primi 5 termini estratti dal corpus italiano con il metodo LLR

Termine lemma	Termine originale
Hq <nsAn	Hqwq Al<nsAn (diritti umani)
>mp mtHd	Al>mm AlmtHdp (Nazioni Unite)
qwp >mn	qwAt Al>mn (forze di sicurezza)
Hryp tEbyr	Hryp AltEbyr (libertà di espressione)
<xtfA' qsry	AlAxtfA' Alqsry (sparizione forzata)

Tabella(15). I primi 5 termini estratti dal corpus arabo con il metodo LLR

4.4.2.1.2. Mutua Informazione (MI)

La seconda misura di *unithood* utilizzata nel nostro studio è la Mutua Informazione (MI). Adottata come strumento standard per misurare l'associazione lessicale nei corpora, la MI confronta la probabilità di dipendenza o coerenza con la probabilità di indipendenza di N-grammi. Nel caso dei bigrammi la formula della MI è come segue:

$$MI(w^1, w^2) = \log_2 \frac{P(w^1, w^2)}{P(w^1)P(w^2)} \quad ,$$

dove il numeratore è la probabilità congiunta di osservare insieme w^1 e w^2 nel corpus, mentre denominatore è il prodotto delle probabilità marginali delle due parole. L'interpretazione della formula precedente è più semplice: maggiore il valore della MI, maggiore è il grado di associazione o coerenza lessicale tra w^1 e w^2. Tuttavia, lo stato dell'arte della MI evidenzia che questa misura è estremamente sensibile ai casi di *hapax*, ovvero ai N-grammi a poca frequenza nel corpus d'acquisizione, il che significa che con questo ultimo tipo di N-grammi il valore di MI risulta evidentemente alto rispetto ad altri N-grammi a frequenza maggiore, i quali potrebbero avere grado di associazione lessicale più forte. Un modo assai comune per attenuare tale difetto della MI è stabilire una soglia di frequenza, che varia a seconda della lunghezza del corpus nonché del tipo di analisi da effettuare[235], in modo che si escludano tutti i N-grammi a frequenza inferiore al valore prestabilito. In effetti, la strategia di adottare soglie di frequenza contiene ancora "lo svantaggio di ridurre drasticamente la quantità di candidati individuati"[236], in quanto un gran numero di termini candidati non saranno considerati semplicemente perché hanno una frequenza rara nel corpus. Nel caso, però, che si affianchi alla MI un'altra misura di associazione meno sensibile ai casi di *hapax*, come LLR, quest'ultima imperfezione si può superare.

Nel caso dei trigrammi o 4-grammi la frazione della MI avrà il seguente formato:

235 Lenci, A. et al., *Testo e computer,* op cit, p.204
236 Ibidem

$$MI(w^1, w^2, w^3) = \log_2 \frac{P(w^1, w^2, w^3)}{P(w^1)P(w^2)P(w^3)}$$

$$MI(w^1, w^2, w^3, w^4) = \log_2 \frac{P(w^1, w^2, w^3, w^4)}{P(w^1)P(w^2)P(w^3)P(w^4)}$$

Le due tabelle seguenti indicano i primi 5 termini delle due liste dei termini selezionati con MI.

Termine lemma	Termine originale
criminalità organizzato	criminalità organizzata
conflitto armato	conflitto armato
equità processuale	equità processuale
Amnesty International	Amnesty International
minoranza etnico	minoranza etnica

Tabella(16). I primi 5 termini estratti dal corpus italiano con il metodo MI

Termine lemma	Termine originale
$gl $Aq	Al>$gAl Al$Aqp (lavori forzati)
Tb $rEy	AlTb Al$rEy (medicina forense)
Hbs AnfrAdy	Hbs AnfrAdy (detenzione in isolamento)
$Ahd EyAn	$hwd EyAn (testimone oculare)
AEqAl tEsfy	AEtqAl tEsfy (arresto arbitrario)

Tabella(17). I primi 5 termini estratti dal corpus arabo con il metodo MI

Dopo il calcolo statistico del grado di associazione delle componenti dei termini candidati, si procede a unificare le due liste per creare un'unica lista dei termini che rappresenta l'input della fase successiva.

```
For candidato in termini candidati:
   LLR_list .append((candidato, LLR_score))
   for term, LLR_score in  LLR_list:
      if  LLR_score > soglia:
         lista_termini_LLR.append((term, LLR_score))
         sorted (lista_termini_LLR, key = LLR_score)

For candidato in termini candidati:
   MI_list .append((candidato, MI_score))
   for term, MI_score in  MI_list:
      if  MI_score > soglia:
         lista_termini_MI.append((term, MI_score))
         sorted (lista_termini_MI, key = MI_score)

for list in  list(lista_termini_LLR, lista_termini_MI):
   values = {}
    for index, term in enumerate(reversed(list)):
       if not term in values:
          values[term]+= index
   return sorted(values.keys(), key = lambada term: values[term]. Reverse = True)
```

Tabella (18). Algoritmo di classificazione dei termini candidati

4.4.2.1.1 Valutazione dei metodi di unithood

Valutare i sistemi di estrazione delle collocazioni in generale e dei termini in particolare rappresenta un'importanza notevole per verificare la performance dei metodi di estrazione anche se il compito di valutazione presenta alcune difficoltà derivate non solo dalla mancanza, in maggior parte dei casi, dei *gold standard* con cui confrontare i risultati ottenuti bensì dall'assenza di una definizione precisa nonché tassativa di *termine* oppure di *collocazione*[237].

237 Pazienza, M.T., et al.,"Terminology extraction", op cit, p.265

Nel nostro caso il processo di valutazione è stato effettuato su una porzione del corpus, i cui risultati ottenuti sono stati poi confrontati con i dati valutati manualmente in termini di *precision, recall* e *f-measure*.

- la *precision* si calcola come la percentuale delle unità lessicali estratte correttamente diviso il numero totale dei termini estratti :

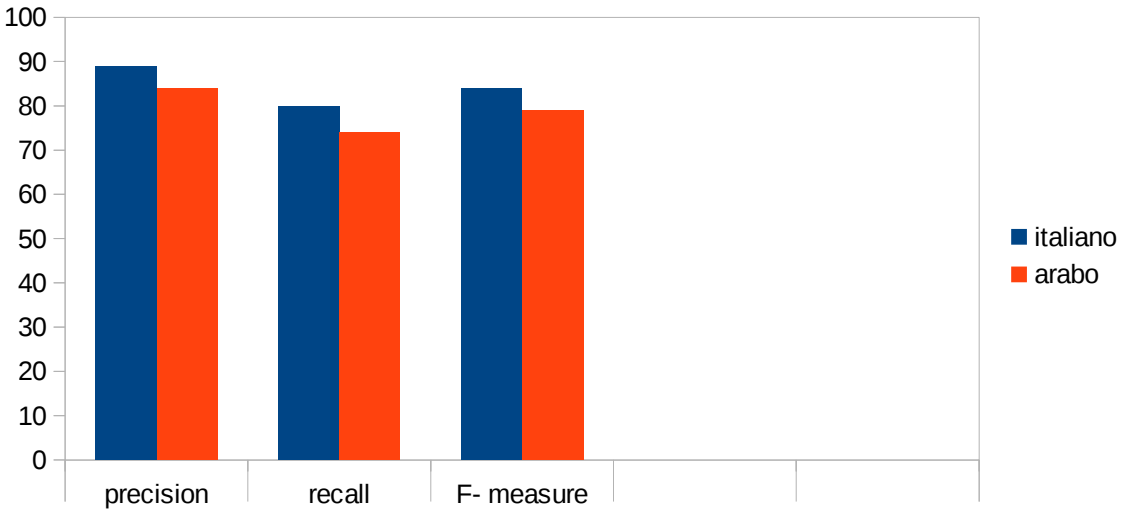
$$precision = \frac{numero\ dei\ termini\ estratti\ correttamente}{totale\ dei\ termini\ estratti} \quad ;$$

- *recall* è definita, invece, come la percentuale delle unità estratte come rilevanti rispetto a tutte le coppie con associazione lessicale nel corpus:

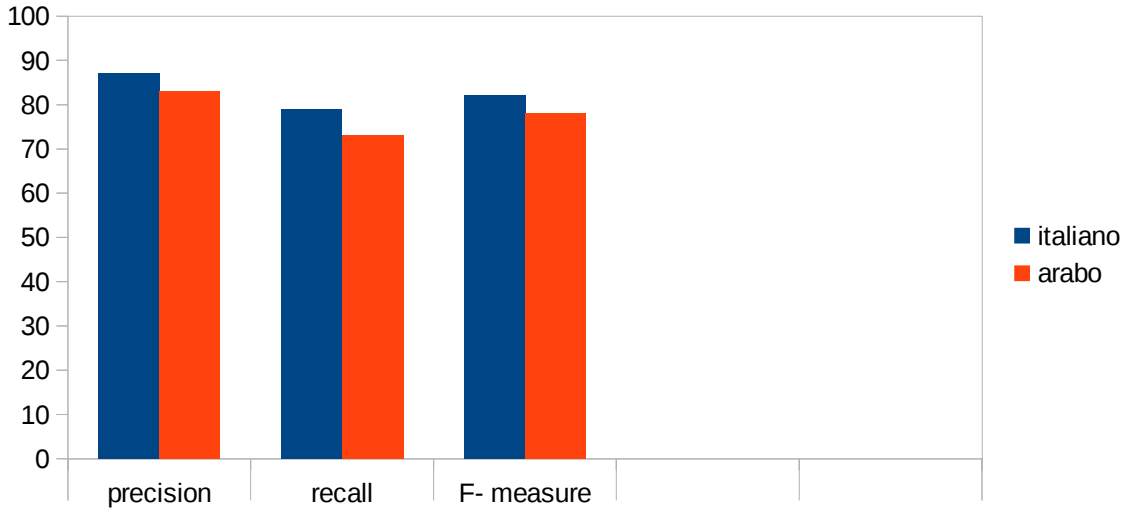
$$recall = \frac{numero\ dei\ termini\ estratti\ come\ rilevanti}{totale\ dei\ termini\ rilevanti\ del\ corpus} \quad ;$$

- $F-measure = 2.\frac{precision\,.\,recall}{precision + recall}$

Le due figure seguenti indicano i dati della valutazione effettuata su entrambe le lingue del corpus. Possiamo notare qualche miglioramento nella performance delle misure con i testi italiani rispetto ai testi arabi, che si può attribuire alla differenza nell'accuratezza nel task di PoS tagging nonché alla natura delle due lingue.



Fig(2). Valutazione del metodo di LLR



Fig(3). Valutazione del metodo di MI

4.4.2.2. Metodi di termthood

4.4.2.2.1. C-NC value

Come è stato presentato precedentemente il metodo *C-NC value* misura la *termhood* di un termine composto integrando le informazioni quantitative di frequenza del termine con quelle contestuali. Nel nostro caso *C-NC value* viene applicata alla lista prodotta dai metodi di *unithood*, cioè LLR e MI, con lo scopo di riclassificarne i termini a seconda della loro *termhood*.

La funzione di *C-value*, che costituisce la prima parte del metodo, si applica alle stringhe più lunghe nella lista di *unithood,* poi scende gradualmente fino ai bigrammi. Le stringhe più lunghe ricevono il loro valore di *C-value* in base alla prima parte della formula, cioè $\log_2(a|)\cdot f(a)$.

In questo caso viene prestabilita una certa soglia di frequenza dei termini, ed i termini sopra questa soglia vengono aggiunti alla lista dei termini candidati. Per le stringhe meno lunghe, per applicare il metodo *C-value* si ha bisogno di ulteriori due informazioni, ossia la frequenza di quelle stringe meno lunghe come parte di termini più lunghi, e il numero di questi termini più lunghi. Per ottenere questi due ultimi informazioni, si procede nella maniera seguente:

per ogni componente *b* del termine composto *a*, estratto come termine candidato, si crea un triplo: *(f(b), t(b), c(b))*,

dove

f(b) è la totale frequenza di *b* nel corpus,

t(b) è la frequenza di *b* come termine annidato, cioè come parte di altri termini più lunghi,

c(b) è il numero di quei termini più lunghi.

Inizialmente, una volta creato questo triplo, *c(b)* = 1, e *t(b)* è eguale alla frequenza di *a*. Successivamente ogni volta viene trovata *b, t(b)* e *c(b)* vengono aggiornati, mentre *f(b)* rimane invariabile. L'aggiornamento di *c(b)* e *t(b)* succede in questo modo: *c(b)* aumenta di uno per ogni ricorrenza di *b* come parte del termine più lungo *a*, estratto come termine candidato; *t(b)* aumenta con la frequenza del termine più lungo *a* ogni qualvolta *b* risulta annidato. Questo vuol dire che se si vuol computare il *C-value* per la stringa (a), si hanno due opzioni: se della stringa (a) non si possiede il triplo *(f(a), t(a), c(a)),* si applica la formula $\log_2 (|a|) \cdot f(a)$, mentre nel caso contrario si utilizza la seconda parte della formula (), cioè,

$\log_2 (|a|) \cdot \left(f(a) - \frac{1}{p(T_a)} \sum_{b \in T_a} f(b)\right)$ e in questo caso $P(T_a) = c(a)$ *e* $\sum_{b \in T_a} = t(a)$.

Il logaritmo utilizzato per compiere questa funzione è descritto come segue[238].

238 Frantzi K., Ananiadou S., "The C–value / NC Value domain independent method", op cit, p.154

for tutte le stringhe *a* di massima lunghezza

calcolare C -value *(a)* = $\log_2 |a| \cdot f(a)$;

if C- value *(a)* ≥ la soglia

aggiungere *a* alla lista dei risultati

for tutti i componenti *b*

rivedere *t(b);*

rivedere *c(b)*

for tutte le stringhe di lunghezza minore

if *a* appare per la prima volta

c value *(a)* = $\log_2 |a| \cdot f(a)$

else

C -value (a) = $\log_2 |a| \left(f(a) - \frac{1}{c(a)} t(a)\right.$

if C- value (a) ≥ la soglia

aggiungere *a* alla lista dei risultati

for tutti i componenti *b*:

rivedere *t(b);*

rivedere *c(b)*;

Il logaritmo precedente prende in input i termini candidati selezionati dalle misure di *unithood* per creare una lista di termini con il loro rispettivo C-value, come nella tabella seguente:

Termine candidato in forma lemmatizzata	C-value
diritto umano	3682
violazione di diritto umano	1850
forza di sicurezza	1751
Nazione Unito	1466
anno di carcere	771
uso eccessivo di forza	695
difensore di diritto umano	678
agente di polizia	622
crimine di guerra	536
libertà di espressione	519

Tabella (19) I primi 10 termini candidati italiani ordinati secondo il valore di C-value

Termine candidato in forma lemmatizzata	C-value
Hq AnsAn (diritti umani)	3877
qwp >mn (forze di sicurezza)	1117
Snf mEAmlp s} (maltrattamento)	934
AnthAk Hq AnsAn (violazione dei diritti umani)	775
Amp mtHd (Nazioni Unite)	766
Hryp tEbyr (libertà di espressione)	763

Enf Dd mr>p (violenza sulle donne)	714
Eqwbp AEdAm (pena di morte)	617
TAlb ljw' (richiedenti asilo)	459
AxtfA' qsry (sparizione forzata)	434

Tabella(20) I primi 10 termini candidati arabi ordinati secondo il valore di C-value

La lista dei termini di C-value rappresentano poi l'input della parte successiva del metodo di *termhood*, cioè NC-value. Come è stato detto precedentemente la funzione NC-value combina il valore di C-value, o qualsiasi altro metodo di *unithood*, di un termine candidato con le sue informazioni contestuali, cioè le parole che appaiono vicine nel testo. Questo metodo comprende tre fasi:

1- applicare la funzione C- value, come è già stato spiegato, per ottenere una lista di termini candidati ordinati a seconda del valore C-value;

2- utilizzare le parole di C- value con valore alto per estrarre le parole di contesto (term context words) a cui assegnare pesi statistici a seconda del numero dei termini con cui appaiono nel corpus;

3- servirsi delle informazioni contestuali fornite nella fase precedente per riclassificare i termini candidati del C-value, aumentando lo spazio tra i termini rilevanti e quelli meno rilevanti. Questo vuol dire che con la funzione NC-value i termini considerati significativi nella lista C-value salgono ancora nella lista, lasciando i posti bassi nella lista ai termini candidati non pertinenti.

Il processo di riclassificazione funziona in questo modo[239]: dalle parole di contesto accompagnanti i termini candidati della lista C-value si considerano solamente i tipi NOMI, AGGETTIVI, VERBI. A ognuna delle parole di contesto viene assegnato un valore statistico in base al numero dei termini che precede o segue nel corpus. Per ogni termine candidato il fattore

239 Ivi, p.170

contestuale viene calcolato sommando il valore statistico delle sue parole contestuali moltiplicato per il numero di occorrenza di queste parole con il termine candidato stesso nel corpus, come nella formula seguente, introdotta precedentemente: *N-value(a)* = $\sum_{w \in C_a} f_a(w) * weitht(w)$.

Combinato, poi, con C-value, questo metodo contestuale dà luogo alla seguente formula: NC-value(a) = 0.5 ∗ C-value(a)+0.5 $\sum_{w \in C_a} f_a(w) * weitht(w)$ [240]

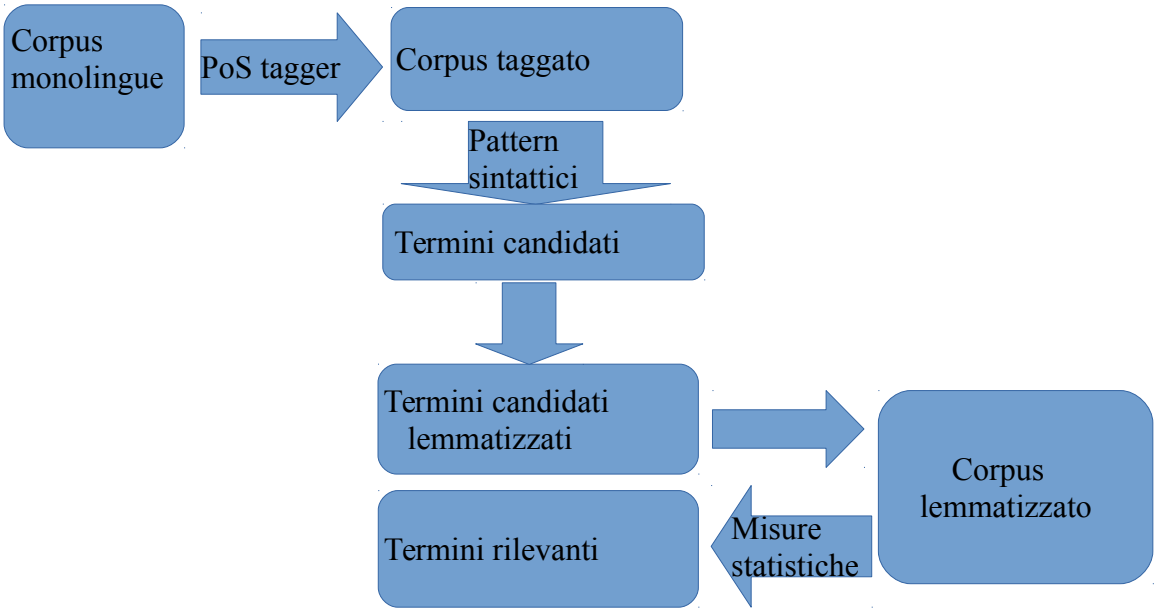
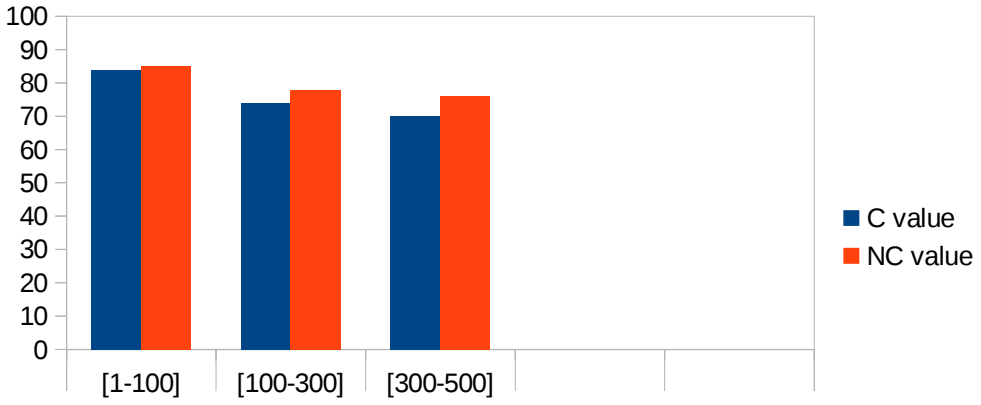


Fig (4). Architettura dell'estrazione monolingue dei termini

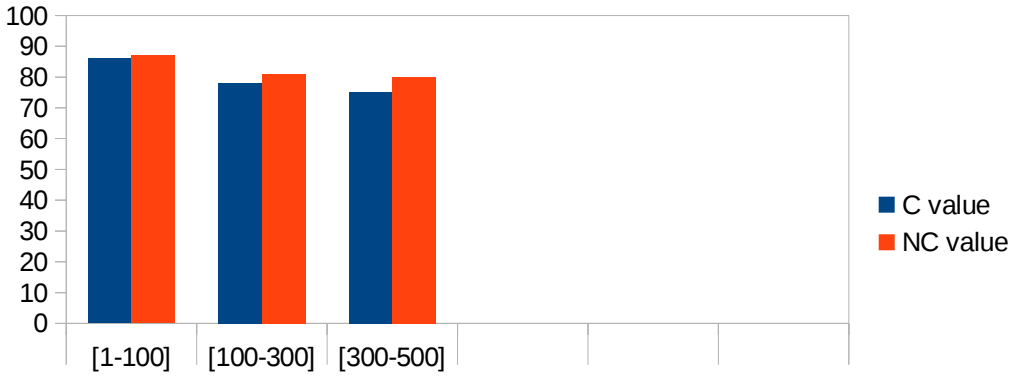
Come dimostrano i diagrammi della precisione dei due metodi (fig(5), fig(6)); la differenza dell'accuratezza nell'identificare termini dal corpus comincia a ampliarsi pian piano che si discende nella lista ordinata dei termini: per i primi 100 termini estratti la differenza di percentuale è del 1% sia per i testi arabi che per quelli italiani; per i successivi 200 termini la

240 Il valore 0.5 viene aggiunto qui solo per dare ottimale distribuzione nel test di precisione; Cfr. Frantzi K., Ananiadou S., op cit, p.171

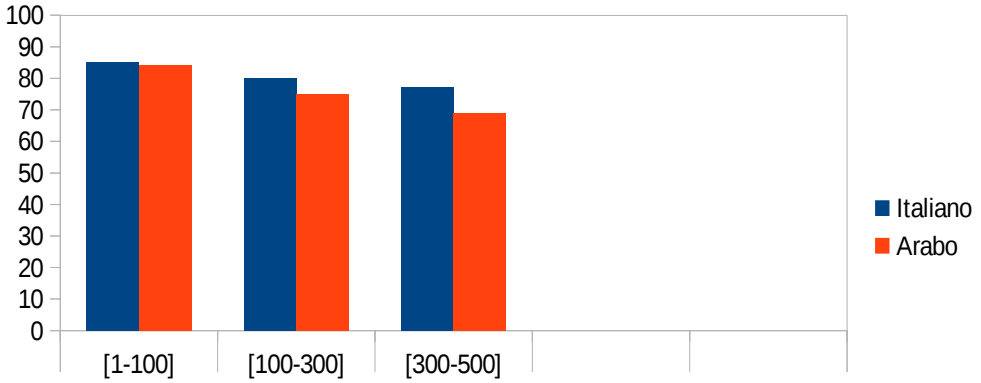
differenza arriva al 4% per l'arabo e al 3% per l'italiano. Di più, per i secondi successivi 200 termini nella lista, cioè dal 300- 500, il divario si avvicina al 6% per i testi arabi e al 5% per il corpus italiano.



Fig(5): differenza in precisione tra C-value e NC value nei testi arabi



Fig(6): differenza in precisione tra C-value e NC value nei testi italiani



Fig(7). Precisione del metodo C-NC value applicato sull'output delle misure di unithood con n-best = 100, 300, 500

Capitolo V: Estrazione dei termini bilingui

5.1. Estrazione di termini da corpora paralleli

Estrarre lessici bilingui o corrispondenze traduttive da corpora paralleli si basa principalmente sulle tecniche di allineamento di due o più testi a livello di parola, che è un compito relativamente complicato rispetto all'allineamento di testi a livello di frase o paragrafo. In effetti, il successo realizzato nel campo di allineare unità testuali a livello di frase (Brown et al.[241]), ha indirizzato gli sforzi dei ricercatori verso la creazione di corrispondenze tra le parole di un corpus parallelo in modo che si possano estrarre unità lessicali bilingui utili in molte applicazioni del TAL come la traduzione automatica, gli studi di lessicologia, la creazione delle ontologie, ecc..

Le sfide di questo compito variano in base al tipo di unità lessicale da allineare, ovvero singole parole o espressioni composte, all'obiettivo dell'allineamento, cioè allineare tutte le parole del corpus parallelo oppure solo alcuni termini o espressioni, nonché alle lingue del corpus, il che significa l'opzione fra lingue che possiedono caratteristiche linguistiche abbastanza comuni (come nel caso delle lingue europee) o lingue che appartengono a famiglie linguistiche diverse (per es. l'inglese e il cinese, il giapponese e l'arabo, ecc.).

La maggior parte dei sistemi finalizzati a estrarre corrispondenze[242] terminologiche si impernia sulle distribuzioni statistiche delle singole unità lessicali all'interno del corpus parallelo partendo dall'assunzione che *translation words are comparably distributed in parallel texts*[243]. Il principio

241 Brown, P., et al., "A statistical approach to machine translation". In *Computational Linguistics*, 16(2):1990, MIT Press Cambridge, MA, USA

242 Gale&Church distinguono in questo caso tra i due termini *allineamento* e *corrispondenza*: mentre il primo si utilizza quando l'ordine delle componenti di due unità (parola o testo) è preservato, il secondo può permettere qualche attraversamento dei rapporti di dipendenza (*crossing dipendencies*), e in quanto nel caso di termini risulta importante effettuare talvolta il *crossing dipendencies,* mentre non lo è, almeno per gran parte delle lingue umane, nel caso di frasi, appare opportuno servirsi del termine *allineamento* parlando di frasi e *corrispondenza* trattandosi di termini o di parole; Gale, W.A., Church, W.K. "Identifying word correspondences in parallel texts". In *Proceedings of the DARPA Speech and Natural Language Workshop* , Pacific Grove, CA, 1991, p.152

243 Teserra, S.A., *Bilingual word and chunk alignment : a hybrid system for Amharic and English*, tesi di dottorato, Bielefeld University, 2007, p.27

di base di tali metodi statistici si incentra sul considerare i dati quantitativi, osservati o stimati, sulle parole del corpus come un indizio significativo per intuire relazioni di traduzione tra di loro. Secondo Fung[244] gli algoritmi probabilistici adottati nell'estrazione di termini o lessici da testi paralleli presentano risultati più affidabili nel caso dei testi specialistici rispetto ai corpora generici, avvalendosi delle seguenti caratteristiche linguistiche dei testi di dominio:
- i termini possiedono un unico significato lungo il corpus testuale;
- i termini hanno un singolo corrispettivo nell'altra lingua del corpus parallelo;
- nella lingua di arrivo non risultano mancanti equivalenti ai termini presenti nella lingua di partenza;
- le frequenze di occorrenza dei termini bilingui sembrano comparabili;
- le posizioni di occorrenza dei termini bilingui risultano, di solito, abbastanza confrontabili.

Nel senso probabilistico si può distinguere tra due approcci adottati per allineare testi a livello di parola in generale e per estrarre lessici bilingui da corpora paralleli in particolare[245]: approcci di associazione (*association approaches*) e approcci di stima (Estimating approach)

5.1. 1. Approcci di associazione

Si tratta di approcci che si basano su un metodo generativo per creare, tramite i diversi metodi statistici che misurano l'associazione lessicale, una lista di equivalenze traduttive candidate a diventare, in base alla probabilità di associazione, delle vere corrispondenze semantiche. Secondo Tiedemann per

244 Fung. P. "A Statistical view on Bilingual Lexicon Extraction: From Parallel Corpora to non-Parallel Corpora". In *Parallel Text Processing:Alignment and Use of Translation Corpora*, Kluwer Academic Publishers, 2000

245 Tiedemann fa notare che il compito di allineare testi a livello di parola si differenzia da quello di estrarre unità lessicali (termini, collocazioni, ecc.) da corpora paralleli, in quanto in questo ultimo caso si possono trascurare i nessi tra le unità lessicali fuori di interesse come le funzioni grammaticali, oppure alcune relazioni di traduzione non certe, ecc. (Tiedemann, J. *Recycling translations: Extraction of lexical data from parallel corpora and their application in natural language processing corpora*, 2003, tesi di dottorato, Acta Universitatis Upsaliensis, p.12, consultato il 20/02/2016: http://stp.lingfil.uu.se/~joerg/phd/html/)

estrarre unità lessicali da corpora paralleli utilizzando questo approccio si deve passare per le seguenti fasi:
1-segmentazione lessicale: in questa fase si identificano i confini delle voci lessicali in ambedue le lingue;
2. corrispondenza: si creano qui delle possibili relazioni di traduzione tra le parole del corpus a seconda di certi criteri di corrispondenza, al fine di produrre un dizionario di traduzione con dei valori di associazione;
3. allineamento e estrazione: le unità lessicali con valori di associazione più affidabili si allineano per permettere poi l'estrazione di termini o collocazioni bilingui.

Gli algoritmi che adottano questo tipo di approccio seguono di solito questo procedimento[246]:
1. scegliere funzione di similarità *S* tra le parole tipo in *SL* e le parole tipo in *TL*;
2. calcolare il valore di associazione *S*(u, v) per un gruppo di coppie di parole tipo (u, v) ∈ (*SL* × *TL*) che occorre nel corpus di addestramento;
3. ordinare le coppie in modo discendente a seconda del valore di associazione;
4. trascurare i candidati con un *S*(u, v) sotto il valore soglia. E il resto delle coppie entra a fare parte delle corrispondenze di traduzione.

La funzione di similarità viene spesso rappresentata dal test *Dice coefficient* o dalle sue varianti. Per ogni frase bilingue si crea una matrice dei valori di associazione tra le parole nelle loro posizioni nel contesto bilingue:

$dice(e_i, f_j) = \frac{2.C(e_i, f_j)}{C(e_i).C(f_j)}$, dove il numeratore rappresenta il numero di co-occorrenza di *e* (token in *SL*) e *f* (token in *TL*) nel corpus parallelo, mentre il denominatore indica la frequenza individuale di *e* e *f* rispettivamente. La funzione di allineamento $a_j = \text{argmax}_i \{\text{dice}(e_i, f_j)\}$[247].

246 Melamed, I. D., "Models of translational equivalence among words". In *Computational Linguistics,* 26, 2, 2000
247 Och, F. J., *Statistical Machine Translation: From Single-Word Models to Alignment Templates*. PhD thesis, RWTH Aachen, Germany, 2002, p.28

Tuttavia, i valori di associazione nella fase 2 del procedimento precedente si calcolano, come fa notare Melamed, in modo indipendente l'uno dall'altro, il che conduce alla cosiddetta *associazione indiretta,* cioè una relazione di associazione tra un token in *SL* e un token in *TL* che non sono semanticamente corrispondenti ma acquistano un valore alto di associazione grazie alla co-occorrenza di uno di loro con il corrispettivo dell'altro. Come soluzione a questo problema Melamed ha proposto il suo metodo *competitive linking algorithm* basato sull'assunzione di allineamento del tipo *one-to-one,* che in una prima fase classifica in ordine discendente i valori di associazione della matrice, poi allinea i token con maggiore valore e in una fase successiva elimina dalla matrice le loro righe e colonne corrispettive.

The competitive linking algorithm can be viewed as a heuristic search for the most likely assignment in the space of all possible assignments. The heuristic is that the most likely assignments contain links that are individually the most likely. The search proceeds by a process of elimination. In the first search iteration, all the assignments that do not contain the most likely link are discarded. In the second iteration, all the assignments that do not contain the second most likely link are discarded, and so on until only one assignment remains[248].

Un esempio dell'approccio di associazione è in Tufiş[249] dove si è presentato un tentativo di estrarre automaticamente equivalenze di traduzione da corpora paralleli taggati a livello morfo-sintattico. L'algoritmo di base, che rappresenta la prima fase dell'approccio, si basa sulle seguenti assunzioni:
1- "1:1 mapping hypothesis", per cui si assume che un token[250] nella lingua di partenza corrisponda a un token anche nella lingua di arrivo;
2. a livello di unità di traduzione (TU), ogni token lessicale, anche se appare polisemico, si utilizza con lo stesso significato;

248 Melamed, D., *Empirical Methods for Exploiting Parallel Texts ,* op cit, p.233
249 Tufiş, D., Barbu, A.M."Lexical token alignment: experiments, results and applications". In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA), 2002
250 Si noti qui l'uso conveniente del termine *token* e non *parola,* in quanto *token* indica un'unità lessicale con un unico significato, che potrebbe essere costituito da più di una parola, e che viene effettuato normalmente da appositi programmi, chiamati segmenter o tokenizer, mentre il termine *parola* indica ogni segno grafico nel testo scritto. Cfr. Ivi, p.459

3. un token in una parte della TU può essere allineato a un token nell'altra parte della TU solo se entrambi i token hanno un PoS tagging compatibile. La compatibilità significa, in questo caso, o ambedue i token possiedono lo stesso PoS, oppure, data una prestabilita definizione, un certo PoS in una lingua ha come corrispondente un altro PoS nella seconda lingua, come per es. quando un gerundio in una lingua si trasforma in un nome in altre lingue;
4. malgrado l'ordine delle parole nella frase non sia un fatto invariante, ci si può concentrare sui token che ricorrono in posizioni relativamente vicine a livello di frase.

In base al tipo di PoS tagging si crea inizialmente una lista candidata di equivalenze di traduzione, in cui ogni tipo PoS contiene le diverse paia di token ($token_s$, $token_t$) che corrispondono al PoS nella stessa TU. Quindi se si hanno la TU^j e la PoS_k, con la raccolta di tutti i token che corrispondono alla PoS_k nelle due parti del Tu^j (nello stesso ordine in cui appaiono nel testo parallelo e dopo la rimozione dei duplicati), si creano due set ordinati: L^{Sj} PoS_k e L^{Tj} PoS_k. Quindi per ogni PoS_i $TU^j{}_{PoSi}$ = L^{Sj} PoS_i ⊗ L^{Tj} PoS_i. Ne consegue che le connessioni del TU^j = CTU^j = $\bigcup_{i=1}^{n.pos} TU^j_{POS_i}$, e quindi in un *n* unità di traduzione in un corpus parallelo la lista candidata delle equivalenze di traduzione = $\bigcup_{i=1}^{n} CTU^j$.
Questa lista viene poi filtrata, in un modo iterativo, con l'uso dei test di associazione lessicale, come il likelihood ratio test con un valore soglia uguale a 9.
Segue l'algoritmo di base un ulteriore algoritmo (Beta) finalizzato a migliorare i risultati e soprattutto evitare gli errori riscontrabili nella prima fase, e in particolar modo quelli causati dalle associazioni indirette. Nel nuovo algoritmo il valore di associazione non è considerato in modo globale a livello del testo parallelo, bensì entro ogni unità di traduzione. Inoltre, nel caso che due o più equivalenze di traduzione abbiano in comune, entro la stessa TU, un token, e ricevano nel frattempo un valore di associazione uguale, la decisione sarà affidata a uno dei due test euristici: similitudine tra le stringhe, e la distanza relativa dei token nel corpus.

Oltre ai mezzi di associazione lessicale, ci sono pure quei metodi che misurano il grado di similarità delle stringhe. L'idea di queste misure parte dall'ipotesi che le parole affini (*cognates*) in due o più lingue etimologicamente legate possano essere buone candidate a rappresentare delle corrispondenze traduttive, e che agli algoritmi che misurano la corrispondenza delle stringhe si possa affidare il compito di identificare tali parole o termini. Per esempio in Brew & McKelvie[251] si utilizzavano varianti di Dice's Coefficient per estrarre parole ortograficamente simili.

Nel suo lavoro, Tiedemann ricorda che accanto alle misure statistiche si possono utilizzare altri strumenti per identificare lessici bilingui o almeno per migliorare i risultati degli altri mezzi adottati. Questi strumenti, che Tiedemann chiama risorse di allineamento esterne, possono consistere in:
- dizionari bilingui in formato elettronico;
- l'ordine delle parole e le loro relazioni di posizione nelle frasi allineate;
- l'integrazione delle annotazioni morfo-sintattiche, attravero il PoS tagging, o delle funzioni sintattiche, tramite i parser sintattici.

5.1.2. Approcci di stima

Gli approcci di stima adottati al fine di allineare due testi paralleli a livello di parola sono ispirati ai modelli della traduzione automatica statistica che rappresentano un'applicazione del *modello di canale rumoroso* (noisy channel model) utilizzato nella teoria dell'informazione[252]. L'idea di base di questi approcci è di sviluppare un modello di traduzione, cioè la relazione di corrispondenza tra stringhe di una lingua d'origine e stringhe di una lingua d'arrivo, i cui parametri vengono valutati tramite la teoria della stima, considerando come variabile invisibile l'allineamento delle parole[253]. Si tratta quindi di metodi che si basano su "building from data a statistical bitext

251 Brew C., McKelvie, D.,"Word-pair extraction for lexicography". In *Proceedings of International Conference on New Methods in Natural Language Processing*, Bilkent, Turkey, 1996.
252 Cfr. Shannon, C.,"A mathematical theory of communication". In *Bell System Technical Journal*, 27, 1948.
253 Cfr. Vulić I., *Term Alignment. State of the Art Overview* , Katholieke Universiteit Leuven, 2010

model the parameters of which are to be estimated according to a given set of assumptions. The bitext model allows for global maximisation of the translation equivalence relation, considering not individual translation equivalents but sets of translation equivalents (sometimes called assignments)[254].

In Brown et al è stato presentato un approccio statistico per gli esperimenti di traduzione automatica. Data una frase *T* nella lingua di arrivo (TL), il compito del loro metodo è quello di trovare la frase *S* nella lingua di partenza (SL), che è la più probabile di essere responsabile della produzione della frase *T*. Se *t* è una stringa della *T* e *s* è una stringa della *S*, la relazione tra *t* e *s* viene designata tramite la probabilità Pr(t | s) che indica la probabilità che *t* sia la traduzione di *s*. Utilizzando il teorema di Bayes questa probabilità può essere formulata nel modo seguente: $Pr(t|s)=\frac{Pr(t)Pr(s|t)}{Pr(s)}$

Trascurando il denominatore Pr(*s)* nell'equazione precedente, in quanto *s* è indipendente da *t*, e usando l'argomento del massimo (argmax) dei due fattori (*s*|*t*), Pr(*t*), si ha la nuova equazione:

$\hat{S} = \text{argmax}_s \Pr(s|t)\Pr(t) = \text{argmax}_s \Pr(S, T)$[255]

dove Pr(s|t) è la probabilità di traduzione di *s* data la presenza di *t*, cioè un modo per presumere le parole nella SL che potrebbero aver prodotto le parole che osserviamo nella TL, mentre Pr(t) indica la probabilità del modello del linguaggio, cioè un modo per mettere in ordine quelle parole stimabili nella TL.

Nel caso del modello di traduzione Pr(s|t), bisogna aggiungere la variabile casuale *a* che indica la relazione di allineamento tra le parole delle stringhe in SL e in TL. Quindi la likelihood di Pr(s|t) diventa Pr(s, a|t) = $\sum_a Pr(s,a|t)$.

Questo tipo di modello allinea ogni parola nella TL ad una parola nella SL, fornendo accanto ad ogni parola della TL la posizione del corrispondente nella frase della SL. Nel caso non ci sia un corrispondente nella SL, la

254 Cfr. Tufiş, D., Barbu, A.M., "Lexical Token Alignment, op cit, p.458

255 La probabilità congiunta Pr(S, T) è il prodotto della probabilità Pr(S), calcolabile tramite il modello del linguaggio LS, e la probabilità condizionata Pr(T|S), calcolabile tramite il modello di traduzione. Cfr. Peter F. et al. "A statistical approach to machine translation". In *Computational Linguistics* , 16(2):79–85, 1990.

posizione sarà 0. Se la frase S nella SL è composta delle parole $s_1 s_2 \ldots s_m$, e la frase T nella TL è composta delle parole $t_1 t_2 \ldots t_l$, quindi una relazione di allineamento $\mathbf{a} = a^m = a_1\, a_2 \ldots a_m$ con $m \in \{0, .., l\}$. La variabile di allineamento $\mathbf{a} = a_l^m$, dove m e l sono la lunghezza di S e T rispettivamente. Il valore di ogni a_j rappresenta la posizione di t_{aj} a cui corrisponde s_j. Quindi se una parola della SL nella posizione j corrisponde ad una parola della TL nella posizione i, $a_j = i$. E se j non corrisponde a nessuna parola nella TL, in questo caso $a_j = 0$. Così il modello di allineamento si può formulare in questa maniera[256].

$$Pr(s, a|t) = Pr(m|t) \prod_{j=1}^{m} Pr(a_j|a_1^{j-1}, s_1^{j-1}, m, t)\, Pr(s_j|a_1^j, s_1^{j-1}, m, t)$$

Quest'equazione parte dai dati osservati, cioè la stringa t e la sua lunghezza l, per arrivare ai dati stimabili.

Per stimare i parametri dell'equazione precedente bisogna adottare misure di approssimazione utilizzando diverse assunzioni di indipendenza. Questo sarà possibile grazie ai modelli di traduzione automatica che si servono di solito del famoso algoritmo di attesa-massimizzazione (expectation-maximization algorithm (EM)), che serve per massimizzare la likelihood delle stime fornite dal corpus di addestramento. L'algoritmo adotta un percorso iterativo per stimare i parametri "nascosti", che sono in questo caso le probabilità di allineamento, data la presenza dei dati osservati, che sono rappresentati qui dalle frasi allineate del corpus.

EM is useful if there are "hidden" parameters which cannot be estimated directly from data. Alignment probabilities are typical examples of such parameters because links between words are not present in the training data. EM starts with an initial guess for all free parameters in the model and updates them iteratively by maximizing the likelihood function until the process converges at a local maximum.[257]

256 Cfr. Peter F. et al.,"The mathematics of statistical machine translation: Parameter estimation". In *Computational Linguistics*, 19(2):263–311, June 1993. p.270

257 Tiedemann, J. *Recycling translations*, op cit, p.23

Brown et al[258] hanno presentato cinque modelli (IBM) finalizzati ad assegnare una probabilità ad ogni relazione di allineamento delle parole di due frasi allineate a livello di frase. I modelli assumono l'ipotesi che ogni parola nella TL corrisponda a una sola parola nella SL.

In modello 1 si assume che gli allineamenti siano indipendenti l'uno dall'altro, con una distribuzione uniforme, quindi la congiunta likelihood della *s* e della *a*, data una stringa *t* avrà la nuova formula:

$$Pr(s,a|t)=\frac{Pr(m|l)}{(l+1)^m}\prod_{j=1}^{m} Pr(s_j|t_{aj}) \quad .$$

Se nel modello 1 l'ordine delle parole nelle due stringhe *s* e *t* non influenza Pr(s|t), in modello 2 l'allineamento tra gli elementi di due stringhe dipende prima dalla posizione degli stessi elementi nelle stringhe nonché dalla lunghezza di queste ultime. Pertanto si introduce il parametro di posizione, *distortion*[259], al modello di traduzione, il che significa la dipendenza dall'ordine delle parole nelle stringhe. Quindi l'equazione del modello 2 diventerà $Pr(s,a|t)=Pr(m|l)\prod_{j=1}^{m} Pr(a_j|j,l,m)\,Pr(s_j|t_{aj})$.

Nei modelli 3 e 4 si aggiunge il parametro di fertilità, cioè la probabilità che una parola in una lingua possa allinearsi a più di una parola nell'altra lingua. Per identificare questo tipo di relazione si effettua l'inversione del modello di allineamento nonché la dipendenza dagli allineamenti precedenti nella frase e dalle classi di parole. In Modello 5 si cerca di eliminare il fenomeno di deficienza presente nei modelli 3 e 4, che significa la presenza di allineamenti "impossibili", quando cioè nella TL si possono generare posizioni fuori della lunghezza di frase.

Una variante dei modelli di traduzione presentati precedentemente sono i tentativi di utilizzare i modelli nascosti di Markov (*Hidden Markov Model* –

258 Peter F. et al., "The mathematics of statistcal machine translation", op cit.

259 Per distortion si intende che nell'ambito dell'allineamento a livello di parole tra due lingue possiamo trovare che una parola in certa posizione in una lingua si allinea con un'altra parola dell'altra lingua non nella stessa posizione bensì in una posizione diversa nella frase; Cfr. Peter F. et al. "A statistical approach to machine translation", op cit.

HMM)[260]. L'idea principale di questi modelli è che per creare allineamento tra le parole di due frasi allineate le probabilità statistiche di allineamento non devono basarsi principalmente sulle posizioni assolute delle parole nelle frasi, bensì sulla differenza tra queste posizioni. In questo approccio per determinare a_j, cioè l'allineamento tra la parola s_j nella posizione j e la parola t_i in posizione i, si deve considerare pure a_j- l, ovvero l'allineamento precedente a livello di frase, quindi la probabilità di a_j diventerà:
$Pr(a_j|a_j\text{-}l, l)$, dove l è la lunghezza della frase t. Quindi, introducendo gli allineamenti "nascosti", $a_1^J = a_1\ a_2\ ..a_J$ delle due frasi $[\ s_1^J|t_1^l\ ;]$, abbiamo la probabilità di traduzione:

$$Pr(s_1^J|t_1^l) = \sum_{a_1^J} \prod_{j=1}^{J} Pr(s_j, a_j|s_1^{j-1}, a_1^{j-1}, t_1^I) = Pr(a_j|, a_{j-1}, I).Pr(s_j|t_{a_j})$$

Tuttavia, i metodi precedenti non riescono ad affrontare i problemi riguardanti l'allineamento delle repressioni composte, ovvero corrispondenze non solo del tipo one to one, le parole di scarsa occorrenza nel corpus (sparse data), oppure il fenomeno di *word-reordering,* cioè la differenza di posizione tra le parole all'interno delle frasi bilingui, ecc[261].
Per far fronte a queste sfide e ottimizzare, quindi, la performance dell'allineamento sono stati proposti tanti tentativi che pur avendo in comune la necessità di integrare i metodi statistici con strumenti esterni, come dizionari bilingue, caratteristiche linguistiche delle parole bilingui, ecc., differiscono nel modo di applicare tali informazioni supplementari.

In Och[262] si è cercato per esempio di raggruppare, tramite un algoritmo di *clustering*, le parole in classi di equivalenza.
"We define bilingual word clustering as the process of forming corresponding word classes suitable for machine translation purposes for a pair of languages using a parallel training corpus".

260 Vogel, S., et al., "HMM-based word alignment in statistical translation". In *Proceedings of the 16th International Confernece on Computational Linguistics*, Copenhagen, Denmark, 1996
261 Cfr. Schrader. B. *Exploiting Linguistic and Statistical Knowledge in a Text Alignment System*. PhD thesis, Universität Osnabrück, 2007, p.45
262 Och, F., "An Efficient Method for Determining Bilingual Word Classes". In *EACL '99*:Ninth Conf. Of the Europ.Chapter of the assocication for computational linguistics, Bergen, Norway, 1999

Se si hanno due classi di vocaboli *SC* e *TC*, tramite la massimizzazione della probabilità congiunta del corpus parallelo, si può effettuare classificazione delle parole del corpus con la seguente formula:

$$(TC', SC') = arg\,max_{TC,SC}\, p(t_1^I, s_1^J | TC, SC)$$
$$= arg\,max_{TC,SC}\, p(t_1^I | TC).p(s_1^J | t_1^I; TC, SC)$$

dove *TC* è una classe delle parole nella TL, *SC* è una classe nella SL. Il moltiplicando dell'ultima equazione viene ottenuto dalla formula seguente:

$p(w_1^N | C) = \prod_{i=1}^{N} p(C(w_i) | C(w_{i-1})).p(w_i | C(w_i))$, mentre il moltiplicatore, la probabilità di traduzione, si calcola con l'assunzione della relazione di allineamento a_1^J nella stessa maniera come spiegato precedentemente, cioè tramite l'equazione: $p(s_1^J | t_1^I; TC, SC) = \prod_{j=1}^{J} P(SC(s_j) | TC(t_{a_j})).p(s_j | SC(s_j))$

Per riconoscere le unità poliremathiche in un corpus parallelo e, quindi, tokenizzarle come singoli token Tiedemann[263] ha proposto un approccio, dove in una prima fase ci si serve delle statistiche delle parole nel testo (similarità tra le stringhe, calcolo delle frequenze a livello monolingue, nonché i valori di co-occorrenza nel corpus parallelo), poi il sistema utilizza in una seconda fase le informazioni linguistiche delle parole (PoS tagging, parsing sintattico, posizione delle parole all'interno della frase). Le caratteristiche statistiche e linguistiche vengono poi combinate in una forma di *clue* che rappresentano un insieme di valori di associazione utilizzato per determinare relazioni di traduzione tra le parole bilingui. La novità di questo metodo è la possibilità di permettere che una parola in una lingua possa connettersi con più di una parola nell'altra lingua, il che aiuta a riconoscere le collocazioni lessicali nel testo e segmentarle, tramite una *dinamica tokenizzazione*, come unità lessicali singole per migliorare poi la performance dell'allineamento.

Simile è l'approccio di Moore[264] che combina le informazioni statistiche con

263 Tiedemann, J., "Combining clues for word alignment". In *Proceedings of the 10th Conference of the European Chapter of the ACL (EACL)*, Budapest, Hungary, 2003

264 Moore, R. C., "Towards a simple and accurate statistical approach to learning translation relationships

quelle linguistiche per riconoscere le collocazioni lessicali in un corpus parallelo e addestrare poi l'algoritmo su questi dati nuovi al fine di affrontare il problema delle associazioni indirette tra parole che non sono equivalenti.

5.2. Stato dell'arte dell'estrazione di termini bilingui

In Daille et al[265] è stato proposto un metodo per estrarre termini bilingui da un corpus parallelo inglese-francese. In una prima fase si selezionano, a livello monolingue con l'ausilio delle informazioni morfo-sintattiche fornite dall'annotazione a livello PoS tagging, termini composti candidati che poi vengono filtrati tramite le misure statistiche. Il risultato finale di questa prima fase è una lista di termini composti inseriti nel corpus parallelo, in attesa dell'estrazione bilingue che rappresenta la fase successiva. Il principio di base del processo dell'estrazione bilingue parte essenzialmente dall'idea che in un corpus parallelo se un termine candidato rappresenta un vero termine, in tal caso sarà abbastanza facile rinvenire il suo equivalente nell'altra lingua, partendo dall'ipotesi che ogni termine in SL si traduca come termine in TL. Come si vede tale approccio non prende in considerazione le variazioni terminologiche nel processo dell'estrazione bilingue.
L'estrazione bilingue consiste di due passi:
1- inizialmente si usa il metodo di frequenza dei termini candidati nel corpus allineato, che viene definito come "conto bilingue". Ad un termine nella SL si associa un termine nella TL se entrambi presentano un grado di associazione relativamente alto a livello del testo. Accanto alla misura di occorrenza vengono utilizzati anche i pattern morfo-sintattici forniti dal PoS tagging. Tale metodo dipende dal concetto di "affinità di pattern", che indica la probabilità che un pattern nella SL (come per es. NdeN in francese) venga tradotto in un certo pattern nella TL (come per es. NN in inglese).
2- l'altro metodo statistico utilizzato viene applicato non ai termini composti nella SL e nella TL, bensì alle singole unità lessicali che compongono i

among words". In *Proceedings of the ACL workshop on data-driven machine translation*, Toulouse, France, 2001
265 Daille B, et al. "Towards automatic extraction of monolingual and bilingual terminology". In: *Proc 15 th COLING*, Kyoto, Japan, 1994

termini composti nelle due lingue. In questo caso il grado di associazione tra due termini bilingui è la somma dei valori di associazione delle parole che li compongono. L'idea principale di questo ultimo metodo consiste nell'assunzione che se due termini composti risultano come traduzione l'uno dell'altro quindi è molto probabile che le loro componenti abbiano un simile rapporto di traduzione o almeno di associazione.

Negli ultimi anni sono stati presentati tanti approcci finalizzati ad affrontare le peculiarità linguistiche della lingua araba e pertanto a migliorare l'output del suo allineamento, a livello di parola, con altre lingue.

Lahbib et al[266] hanno proposto un metodo di estrarre termini di dominio da corpora allineati arabo-inglese. Questo sistema consiste in: 1) analisi morfologica e disambiguazione delle parole del corpus; 2) estrazione di termini arabi tramite il PoS tagging e il TF-IDF; 3) allineare il corpus a livello di parola utilizzando GIZA++; 4) estrazione di relazioni di traduzione, in base ad una matrice di traduzione generata dal processo di allineamento, che consiste nell'estrarre per ogni parola araba nel corpus la probabile relativa traduzione. Per valutare l'approccio, una versione vocalizzata di hadith corpus[267] è stata usata, con tasso di precisione vicino al 90%.
Riguardo a questo sistema possiamo fare alcune osservazioni: innanzitutto l'approccio si basa su uno strumento probabilistico per allineare i testi a livello di parola. Questo non può garantire buoni risultati trattandosi di lingue come l'arabo che ha le proprie caratteristiche sintattiche e morfologiche. In secondo luogo, il corpus di valutazione è un corpus religioso che contiene terminologie islamiche che non hanno corrispondenti semantici in altre lingue, ma solo traslitterazione.

Partendo dalla teoria che le costruzioni VS in arabo presentano qualche difficoltà quando vengono tradotte automaticamente in inglese, in Marine

266 Lahbib W., et al. "Arabic -English domain terminology extraction from aligned corpora". In *On the Move to Meaningful Internet Systems: OTM* 2014 Conferences. Lecture Notes in Computer Science, Vol. 8841, Springer
267 http://library.islamweb.net/hadith/index.php

Carpuat et al[268] è stato introdotto un metodo per risolvere tale problematica tramite il parser sintattico, ovvero per identificare inizialmente questo tipo di costruzioni e poi riordinarle in costruzione tipo SV, come è il caso in inglese, per migliorare, quindi, la performance dell'allineamento.

In Habash&Sadat[269] sono state introdotte delle tecniche di pretrattamento del testo arabo prima del compito dell'allineamento. Tali tecniche si basano su tre schemi che si utilizzano in un modo graduale. Nel primo schema, REGEX, si usano le espressioni regolari per tokenizzare le unità lessicali separando soprattutto i prefissi o i suffissi che fungono da pronomi clitici. Nello schema BAMA, che è la tecnica successiva, ci si serve dell'analizzatore morfologico Buckwalter per ottenere le diverse analisi delle parole del testo per selezionarne, poi, la prima analisi in testa alla lista. Per disambiguare, invece, le analisi effettuate da BAMA si utilizza il terzo schema, MADA, che, tramite la combinazione di diversi classificatori, determina l'analisi più corretta.

Per fronteggiare il problema della scarsità dei dati causata dalla complessità della morfologia araba, Kfir&Nachun[270] hanno visto che usare equivalenti semantici anziché unità lessicali singole possa contribuire ad aumentare il riconoscimento delle corrispondenze tra il testo di input e il corpus di addestramento, e quindi a migliorare l'efficacia della traduzione automatica tra l'arabo e le altre lingue.

---

268 Carpuat, M. et al., "Improved Arabic-to-English statistical machine translation by reordering post-verbal subjects for word alignment". In: *Proceeding ACLShort '10 Proceedings of the ACL* 2010

269 Habash N. and Sadat F., "Arabic Preprocessing Schemes for Statistical Machine Translation". In *Proc. of the Human Language Technology Conference of the NAACL*, Companion Volume: Short Papers, New York City, NY, 2006.

270 Bar, K., Dershowitz, N., "Using semantic equivalents for Arabic-to-English Example-based translation". In *Challenges for Arabic Machine Translation*, John Benjamins Publishing Company, 2012

5.3. L'approccio della tesi all'estrazione di termini bilingui

La caratteristica principale del nostro approccio all'estrazione di termini bilingui è di non basarsi sugli strumenti di allineamento adottati nei metodi di traduzione automatica. E questo ha le seguenti motivazioni:

- La maggior parte dei sistemi di allineamento tra l'arabo e le altre lingue dipende dai metodi statistici, come GIZA, MGIZA, ecc. i quali come abbiamo visto nella parte precedente si basano sul calcolo delle occorrenze e delle massimizzazioni delle probabilità osservate nel corpus, utilizzando diversi modelli di traduzione automatica. Un tale metodo di allineamento non garantisce, tuttavia, un grado alto di precisione, perché "le corrispondenze rare non saranno considerate, anche se possono essere rilevanti per applicazioni come la terminologia o la lessicologia"[271].

- uno studio, come il nostro, che concerne le variazioni terminologiche non deve dipendere totalmente dai metodi statistici per allineare i termini nel corpus parallelo, dal momento che l'allineamento in questo caso riguarda solo certe unità lessicali e non tutto il testo, il che significa che la mediocre performance realizzabile con i metodi probabilistici avrà sicuramente un impatto negativo sui risultati dello studio. Ne consegue che per garantire una certa recall dei termini bilingui abbiamo deciso di adottare un approccio ibrido che combina le statistiche dei termini con le loro caratteristiche linguistiche nel contesto parallelo.

Partendo dall'ultima fase raggiunta nella parte dell'estrazione monolingue, cioè due liste separate di termini monolingui e un corpus parallelo allineato a livello di frase, l'approccio all'estrazione bilingue segue i seguenti passi:

- restituire i termini estratti nella fase monolingue nel contesto parallelo;
- distinguere, con un formato marcato, i termini inseriti nel corpus parallelo dal resto delle unità lessicali;

271 Ahrenberg L., et al., "Interactive word alignment for language engineering". In *Proc EACL*, Budapest, 2003

- estrarre i termini paralleli a livello di unità di traduzione;
- verificare le corrispondenze traduttive, escludendo i termini candidati che non hanno equivalenti nell'altra lingua;
- creare una lista unica delle corrispondenze di traduzione

La figura seguente illustra i passi dell'approccio

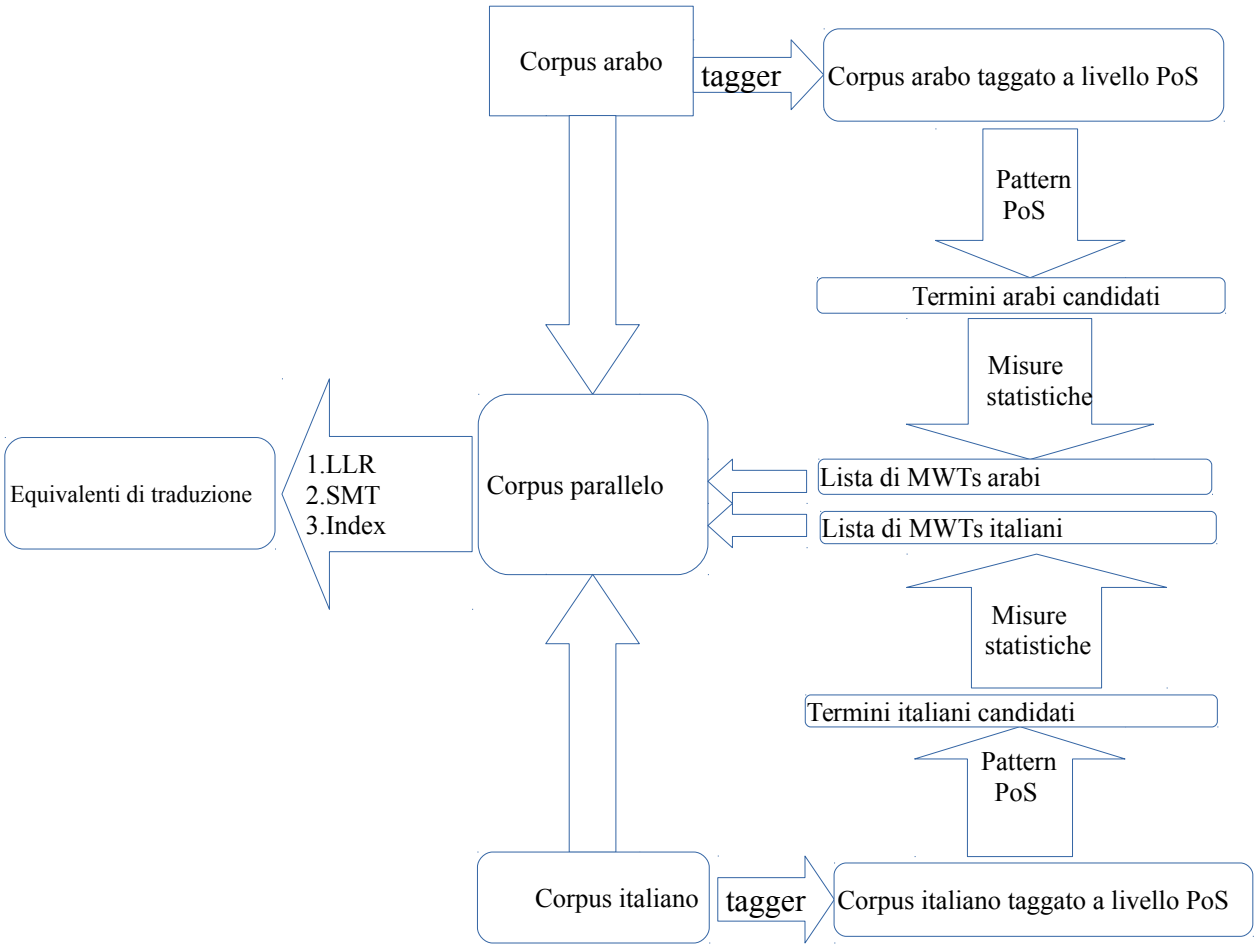


Fig.(8). Architettura del sistema dell'estrazione di termini bilingui adottato nella tesi

5.3.1. Restituire i termini composti nel contesto parallelo

In questo passo i termini monolingui si restituiscono nel contesto parallelo acquisendo un formato marcato, tramite parentesi quadre, rispetto al resto del corpus. Questo viene effettuato tramite un semplice codice come segue:

```
MWTs =  [ w.strip() for w in mwts_file.split(',')] if w != ""]
MW_list = ast.literal_eval(MWTs)
MW = '|'.join(MW_list)
new_file = re.sub('('+MW+')', r'[\1]', parallel_corpus)
print (new_file)
```

Fig.(9): codice della restituzione dei termini nel corpus parallelo

```
<seg match="1-1" id="2">
<src>Difensore di [diritto umano] e organizzazione essere essere oggetto di crescente intimidazione e minaccia in uno clima di limitazione di [libertà di espressione], che avere anche vedere uno giornalista incarcerare per diverso mese.</src>
<tgt>تَعَرَّض عَدَد مِن مُدافِع عَن [حق إنسان] ومنظمات مَعْنِيّ ب حقوق إِنْسان ل ترهيب وتهديد ب صورة متزايد، وَسْط مُناخ مِن قَيْد عَلَى [حرية تعبير]، شَهِد أيضاً حَكَم ب سجن أَحَد صُحُفِيّ ل عدة شهر.</tgt>
</seg>
<seg match="1-1" id="3">
<src>Essere essere riportare caso di [sgombero forzato] e [violazione di diritto umano] da parte di polizia, ma su scala minore rispetto a anno precedente.</src>
<tgt>ووردت نَبَأ عَن حالَة مِن [إجلاء قسري] و[انتهاك حَقّ إنسان] عَلَى يَد شرطة، وإن كان ذلِكَ عَلَى نِطاق أَقَلّ مِن مثيل فِي سَنَة سابقة.</tgt>
</seg>
<seg match="1-1" id="4">
<src>Il [elezione legislativo] e presidenziale rinviare a fine di 2007 essere essere ulteriormente rimandare rispettivo a 2008 e a 2009.</src>
<tgt>وتأجلت مَرَّة أُخْرَى [انتخاب تشريعي] إِلَى عام 2008، وانتخاب رِئاسِيّ إِلَى عام 2009، وكان قَدْ سَبَق تأجيل إِلَى آخِر عام 2007.</tgt>
</seg>
<seg match="1-1" id="5">
<src>Il [agente di polizia] responsabile di tale violazione e per il violazione compiere in 2006 non essere essere assicurare a giustizia</src>
<tgt>ولم يُقدم إِلَى ساحَة عَدالَة [ضابط شرطة] مَسْؤُول عَن هذا انْتِهاك وعن انْتِهاك الَّذِي ارتُكبت فِي عام 2006.</tgt>
</seg>
```

Fig. (10): Il corpus parallelo dopo la restituzione dei termini nel contesto

5.3.2. Estrarre i termini paralleli a livello di unità di traduzione

Successivamente per ogni unità di traduzione del file TMX vengono estratti i termini tra le parentesi quadre. L'obiettivo della presente fase è solo raggruppare i termini bilingui a livello di ogni unità di traduzione. L'output di questo passo è dimostrato nella tabella seguente, dove si vedono i diversi tipi di relazione tra i termini bilingui.

Tipo di relazione	Relazione	Esempi
Relazioni positive	UNO a UNO	sgombero forzato : <jlA' qsry
	UNO a DUE	partito di opposizione : jmEyp wTny, Hzb mEArDp
	UNO a TANTI	violenza sessuale : Enf jnsy, nZAm qDA}y, rEAyp Tby
	DUE a UNO	corte militare, sentenza di morte : Hkm <EdAm
	DUE a DUE	servizio militare, leva militare : tjnyd Eskry, xdmp Eskry
	DUE a TANTI	forza di sicurezza, senza accusa : AlAtHAdyp AlAntqAlyp, qwAt Al>mn, bdwn thmp
	TANTI a UNO	forza di sicurezza, struttura di detenzione, diritto umano : Hq <nsAn
	TANTI a DUE	ufficiale militare, crimine di guerra, corte militare : jrA}mp Hrb, DAbT jy$
	TANTI a TANTI	gruppo armato, arresto di massa, esecuzione extragiudiziale : <EdAm xArj nTAq qDA', Hmlp AEtqAl wAsE, jmAEp mslH
Relazioni negative	UNO a NULLA	rivolta carcerario: []
	DUE a NULLA	società di sicurezza, giornalista freelance: []
	TANTI a NULLA	diritto umano, rapporto sessuale, standard internazionale: []
	NULLA a UNO	[] : jrymp >mn dwl
	NULLA a DUE	[]:Hq <nsAn , n$AT <rhAby
	NULLA a TANTI	[]:jmEyp EAm >mp mtHd , Hkm <EdAm , wqf tnfy*
	NULLA a NULLA	[]:[]

Tabella(21):Relazioni della presenza dei termini bilingui nelle unità di traduzione

5.3.3. Estrarre corrispondenze traduttive

Questa fase si concentra essenzialmente sulle relazioni positive di ogni unità di traduzione e mira a convalidare delle equivalenze traduttive, utilizzando un approccio statistico-linguistico le cui componenti sono le seguenti:

5.3.3.1. Log Likelihood ratio (LLR)

Similmente all'estrazione monolingue ci serviamo anche nell'estrazione bilingue dei calcoli forniti dal LLR test per poter determinare corrispondenze di traduzione tra gli elementi bilingui di ogni unità di traduzione del testo parallelo. Il principio di base in questo passo consiste nel fatto che le statistiche fornite dal corpus parallelo per ogni paio di termini bilingui possono essere utilizzate per creare una tabella di contingenza tramite cui si misura il valore di associazione o co-occorrenza tra questi termini. L'esempio seguente spiega il procedimento: per la coppia bilingue: *sgombero forzato > <jlA' qsry,* possiamo calcolare la probabilità dell'associazione traduttiva della coppia nel contesto parallelo, ovvero la co-occorrenza dei due termini ("sgombero forzato" e "<jlA' qsry") nella stessa unità di traduzione, e la probabilità marginale di ogni parte nel corpus parallelo. Successivamente si applica la formula del LLR test presentata precedentemente per stimare un valore di associazione di traduzione per ogni paio dei termini composti raccolti nella fase precedente.

	sgombero forzato	# sgombero forzato
<jlA' qsry	569	27
# <jlA' qsry	17	59881

Tabella(22): Esempio di una tabella di contingenza dei termini bilingui

Per calcolare il valore LLR di ogni unità di traduzione ci troviamo innanzi a due situazioni: a) calcolare il LLR per unità di traduzione del tipo UNO a UNO; b) calcolare il LLR per unità di traduzione che contengono relazioni multiple. Nel primo caso, cioè nel caso UNO a UNO, si accettano le coppie con valore LLR relativamente significativo. Nel caso delle relazioni multiple la situazione presenta, invece, qualche complessità: inizialmente si effettua un'iterazione tra i diversi termini bilingui, poi si calcola il valore LLR per ogni coppia e, infine, si seleziona l'abbinamento con un valore alto rispettivamente alle altre paia associate che condividano un termine sia arabo che italiano. Ad esempio per l'unità seguente:

libertà di espressione, crimine di guerra, sparizione forzata	jrymp Hrb , AxtfA' qsry, Hryp tEbyr

si procede nella maniera seguente per identificare valide corrispondenze traduttive:

1. creare iterazione tra le parti della lingua di partenza e quelle della lingua di arrivo. Gli abbinamenti prodotti avranno il formato seguente:

('sparizione forzata', 'jrymp Hrb ')
('crimine di guerra', 'jrymp Hrb ')
('libertà di espressione', 'jrymp Hrb ')
('sparizione forzata', 'Hryp tEbyr')
('crimine di guerra', 'Hryp tEbyr')
('libertà di espressione', 'Hryp tEbyr')
('sparizione forzata', 'AxtfA' qsry')
('crimine di guerra', 'AxtfA' qsry')
('libertà di espressione', 'AxtfA' qsry')

2. misurare il valore LLR per ogni paio delle coppie precedenti.

3. fra le coppie che condividono un termine, come le prime tre, o le seconde tre o le terze tre, viene selezionata la coppia con valore LLR alto.

Quindi se applichiamo il LLR test a questo esempio, otteniamo le seguenti statistiche:

[(sparizione forzata, jrymp Hrb): 1594
(sparizione forzata, Hryp tEbyr) :2.3
(sparizione forzata, AxtfA' qsry):1919;

(crimine di guerra, jrymp Hrb):2939
(crimine di guerra, Hryp tEbyr):1188
(crimine di guerra, AxtfA' qsry):1077;

(libertà di espressione, jrymp Hrb): 7275
(libertà di espressione, Hryp tEbyr): 9301
(libertà di espressione, AxtfA' qsry):8205]

Applicando il confronto tra le coppie che hanno in comune un termine composto si ottiene il seguente risultato che corrisponde alle equivalenze di traduzione:
(sparizione forzata, AxtfA' qsry):1919
(crimine di guerra, jrymp Hrb):2939
(libertà di espressione, Hryp tEbyr): 9301

```
corpus_parallelo = file del corpus parallelo lemmatizzato in entrambe le lingue
coppie_file = file che contiene le coppie di traduzione estratte da ogni unità di traduzione e ordinate a seconda del tipo di relazione di traduzione
num =0 # numero dei N-grammi nel corpus
score = 0.1 # per evitare la divisione per il valore 0
for line in corpus_parallelo.splitlines():
    num += 1
for line in coppie_file.splitlines():
    coppie= []
    LLR_list  = []
```

```
    parallelo = line.split(' > ')
    it =  parallelo[0].split(', ')
    ar = parallelo[1].split(', ')
    for item in it:
        for item_ in ar:
            it_ar = it_not_ar = ar_not_it  = 0
            for source_language, target_language in corpus_parallelo:
                if item in source_language and item_ in target_language:
                    it_ar += 1
                elif item in source_language and not item_ in target_language:
                    it_not_ar += 1
                elif item not in source_language and item_ in target_language:
                    ar_not_it += 1
        LLR = log((( it_ar + score)/num)/
(((it_not_ar + score)/num)*((ar_not_it + score)/num)))
            coppie.append((item, item_, LLR))

            for key1, key2 in groupby(coppie, lambda coppie: coppie[0]):
                LLR_list.append(list(key2))
            for key1, key2 in groupby(coppie, lambda coppie: coppie[1]):
                LLR_list.append(list(key2))
            for coppia in LLR_list:
                coppia = sorted(coppia, key=operator.itemgetter(2))
```

Fig.(11): l'algoritmo di estrazione bilingue tramite il LLR test

Termini bilingui	Valore LLR
libertà di espressione >> Hryp tEbyr	9301
consiglio di sicurezza>> mjls >mn	4856
conflitto armato >> nzAE mslH	2602

diritto umano >> Hq <nsAn	2939
corte costituzionale>>mHkmp dstwry	2554
codice penale>> qAnwn Eqwbp	1787
sgombero forzato >> <jlA' qsry	1136
Amnesty International >> mnZmp Efw dwly	4381
pena di morte >> Eqwbp <EdAm	2470
disegno di legge>> m$rwE qAnwn	1057
arresto domiciliare >> <qAmp jbry	1368
agente di polizia >> DAbT $rTp	2932
forza di sicurezza >> qwp >mn	1303
prigioniero di coscienza >> sjyn r>y	2144
richiedente asilo >> TAlb ljw'	4102
procuratore generale >> nA}b EAm	1259
violenza sessuale >> Enf jnsy	2867
stato di emergenza >> HAlp TwAr}	1317
testimone oculare>>$Ahd EyAn	625
standard internazionale>> mEyAr dwly	1295

Tabella(23): Esempi di MWTs bilingui con il loro LLR valore

5.3.3.2. Sistema di traduzione automatica statistica (SMT)

Si utilizza in questa fase Google Translate come un sistema di traduzione automatica: l'idea qui è che tramite la traduzione delle componenti dei MWTs si può identificare termini bilingui equivalenti. Esponiamo l'esempio seguente:

- violazione di diritto umano, caso di sgombero forzato >> AnthAk Hq <nsAn, HAlp <jlA' qsry
- sentenza di morte >> Hkm <EdAm, mHkmp Eskry
- crimine di guerra, corte militare, giudicato colpevole >> jrymp Hrb
- codice penale, prigioniero di coscienza >> sjyn r>y

Nel primo caso la traduzione di Google Translate dei termini italiani appare così:AnthAk Hq mn Hqwq Al<nsAn, HAlp Al<xlA' Alqsry. Dopo l'eliminazione delle preposizioni e la conversione nella forma lemmatizzata, che è la forma dei termini nella lista selezionata, otteniamo AnthAk Hq <nsAn, HAlp <xlA' qsry. Confrontando, in una fase successiva, i risultati di SMT con quelli della lista, possiamo scegliere le seguenti coppie di MWTs come valide traduzioni:

-violazione del diritto umano: AnthAk Hq <nsAn
-caso di sgombero forzato : HAlp <jlA' qsry

Seguendo lo stesso approccio con gli altri due casi, si hanno questi risultati:

- sentenza di morte : Hkm <EdAm
- crimine di guerra : jrymp Hrb
- prigioniero di coscienza : sjyn r>y

Come si può notare dagli esempi, il metodo combina i MWTs la cui parte araba condivida maggior numero di parole con la traduzione SMT della parte italiana.

I motivi principali di servirci di un sistema di traduzione automatica in questa parte del lavoro sono i seguenti: a) per le lingue del nostro corpus non abbiamo trovato risorse lessicali esterne consistenti in dizionari, tesauri, o ontologie nel dominio giuridico; b) i sistemi di SMT forniscono traduzione di alta qualità quando si tratta di singole parole o semplici frasi, mentre la performance della traduzione diminuisce quando le strutture del testo da tradurre cominciano a complicarsi, sia sintatticamente che pragmaticamente[272]; c) la dipendenza da un sistema SMT non rappresenta l'approccio essenziale in questa fase del lavoro, bensì è solo una parte complementare agli atri metodi utilizzati per l'allineamento dei termini a livello bilingue; d) l'utilizzo del sistema SMT serve nel nostro caso solo a selezionare, da una lista già creata, le corrispondenze traduttive corrette, il che vuol dire che qui non si tratta di produrre nuove traduzioni in cui possono avvenire gli errori dovuti in genere alla traduzione automatica.

Le difficoltà principali di tal metodo derivano dalla differenza tra la traduzione SMT e quella presente nel corpus, soprattutto trattandosi di termini giuridici che presentano certe peculiarità a livello lessicale, motivo per cui può succedere che il sistema non combina due coppie bilingui solo perché le unità lessicali prodotte da SMT sono differenti, come nei casi seguenti:

(matrice razzista : dwAfE EnSryp);
(oltraggio a pubblico autorità :AltEdy ElY ms&wl EAm),

dove la traduzione SMT dei termini italiani risulta così:

(matrice razzista : Erqy);
(oltraggio a pubblico autorità :<hAnp AlslTp AlEAmp)

272 Cfr Li H., et al., "Comparison of Google Translation with Human Translation". In *Proceedings of the Twenty-Seventh International Florida Artificial Intelligence Research Society Conference,* 2014

```
For line in coppie_file:
   termini_bi = []
   ar, it = line.split(“>”)
   if la relazione è del tipo UNO a UNO
            if la traduzione_it di ar[273] == it or any(i in it.split() for i in
traduzione_it.split()):
          termini_bi.append(it, ar)

   if la relazione è del tipo TANTI a UNO
      for item_it in it:
           if la traduzione_it di ar == item_it or any(i in item_it.split() for i in
traduzione_it.split()):
            termini_bi.append(item_it, ar)

   if la relazione è del tipo UNO a TANTI
      for item_ar in ar:
          if la traduzione_ar di it == item_ar or any(i in item_ar.split() for i in
traduzione_ar.split()):
            termini_bi.append(it, item_ar)

   if la relazione è del tipo TANTI a TANTI
      for item_it in it:
         for item_ar in ar
             if la traduzione_it di item_ar == item_it or any(i in item_it.split()
for i in traduzione_it.split())
               termini_bi.append(item_it, item_ar)
            if la traduzione_ar di item_it == item_ar or any(i in item_ar.split()
for i in traduzione_ar.split())
               termini_bi.append(item_it, item_ar)
```

Fig.(12): l’algoritmo di estrazione bilingue tramite il SMT

273 Per questo scopo abbiamo utilizzato la versione 1.5.1 di Goslate, come free Google Translate API: https://pypi.python.org/pypi/goslate, utilizzato il 4/03/2015

5.3.3.3. posizione delle parole all'interno delle frasi

In questo passo ci si serve della posizione dei MWTs all'interno della frase nel corpus parallelo. Entro ogni unità di traduzione, il sistema seleziona i MWTs vicini a livello del loro posto nella frase, con una soglia di distanza = 4. In una prima fase vengono estratti i termini composti bilingui a livello di frase accompagnati dal valore che indica la relativa posizione all'interno della frase parallela; poi si selezionano quelli vicini. Nonostante che l'italiano e l'arabo abbiano un sistema libero delle parole a livello di frase, questo non nega il fatto che le loro parole, specialmente le unità poliremathiche, presentino una certa probabilità di avvicinamento a livello di posizione nella frase. Pensiamo ai casi seguenti tratti dal corpus parallelo originale:

1.Sono stati riportati casi di **sgomberi forzati** [5] e **violazioni dei diritti umani** [7] da parte della polizia, ma su scala minore rispetto agli anni precedenti .
وردت أنباء عن حالات من **الإجلاء القسري**[5] و **انتهاكات حقوق الإنسان**[7] على أيدي الشرطة ، وإن كان ذلك على نطاق أقل من مثيله في السنوات السابقة.

2.L' uguaglianza , la collaborazione tra donne e uomini e il rispetto per la **dignità umana** [14] devono permeare tutti i livelli del **processo di socializzazione** [21].
ولا بد أن تشيع المساواة و المشاركة بين المرأة و الرجل ، و احترام **كرامة الإنسان**[15] ، في جميع مراحل **التنشئة الاجتماعية**[20].
3.Lo sfruttamento delle donne nella **prostituzione internazionale** [5] e i canali del **traffico clandestino di donne** [10] sono divenuti vitali per il **crimine internazionale organizzato** [16].
و قد أصبح استغلال المرأة في **الشبكات الدولية للبغاء**[6] و **الاتجار بالمرأة** [8] محور اهتمام رئيسي ل **الجريمة الدولية المنظمة** [13].

Se osserviamo il valore tra le parentesi quadre che indica appunto la posizione dei termini in grassetto (si prende il valore del primo elemento del termine composto) nella frase parallela, possiamo notare evidentemente che la

differenza tra i valori dei termini bilingui rimane spesso nell'ambito del valore 4.

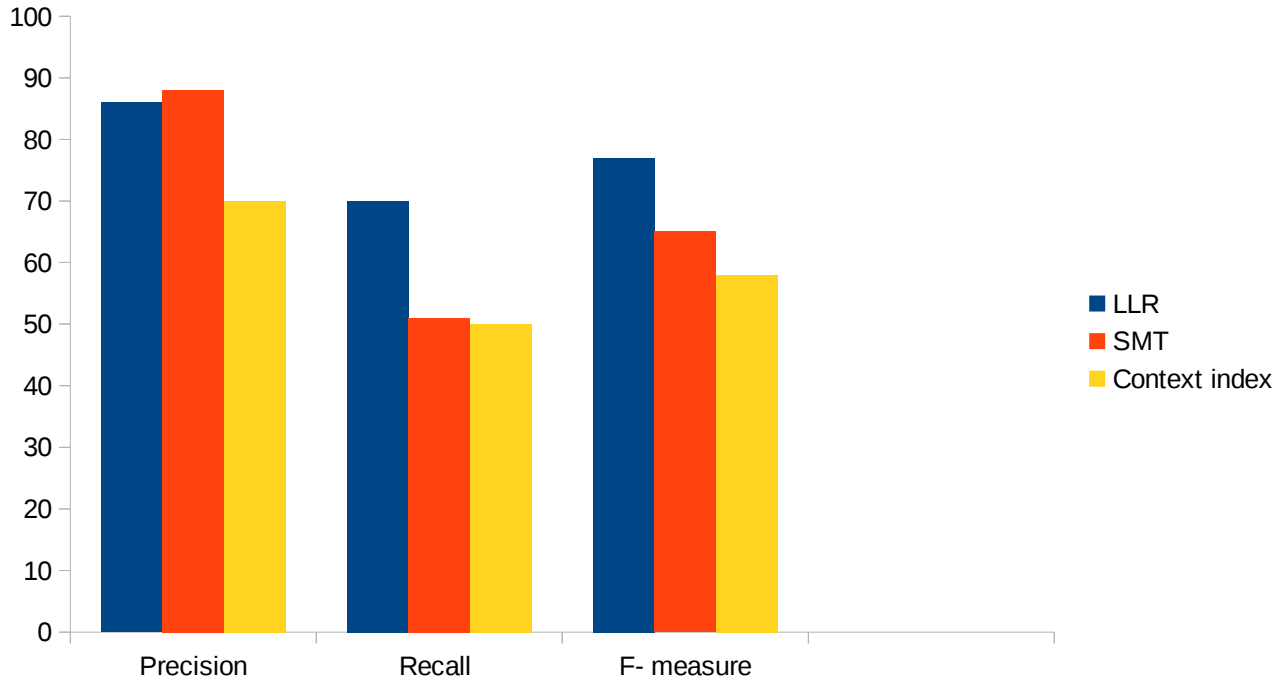
Applicando quindi il metodo di distanza tra le parole ai tre esempi precedenti, otteniamo i risultati seguenti:

sgomberi forzati [5] > Al<jlA' Alqsry[5]
violazioni dei diritti umani [7] > AnthAkAt Hqwq Al<nsAn[7]
dignità umana [14] > krAmp Al<nsAn[15]
processo di socializzazione [21] > Altn$}p AlAjtmAEyp[20]
prostituzione internazionale [5] > Al$bkAt Aldwlyp llbgA'[6]
traffico clandestino di donne [10] > AlAtjAr bAlmr>p [8]
crimine internazionale organizzato [16]. > Aljrymp Aldwlyp AlmnZmp [13]

```
for line in coppie_file# file che contiene i MWTs paralleli separati da '>'
arabo, italiano = line.split(">")
arabo_valore =[int(valore) for valore in re.findall('(\d+)', arabo)]
italiano_valore =[int(valore) for valore in re.findall('(\d+)', italiano)]
arabo = arabo.split(",")
italiano = italiano.split(",")
new_coppie = ([(termine_ar, termine_it) for termine_ar, termine_it in zip(italiano, arabo)] if [abs(it-ar) < 4 for it,ar in zip(italiano_valore, arabo_valore)])
```

Fig.(13): l'algoritmo di estrazione bilingue con l'utilizzo di posizione delle parole

Come dimostra il diagramma nella Fig(13) il processo dell'estrazione bilingue adottato nella tesi presenta una performance relativamente alta nel caso del metodo LLR. Inoltre, il basso *F- measure* nell'estrarre con SMT deriva probabilmente dalla particolarità dei testi giuridici le cui terminologie si presentano spesso diverse dalla traduzione di Google Translate; mentre il mediocre *F-measure* con il metodo basato sulla posizione dei termini all'interno della frase si attribuisce all'ordine libero delle parole nella frase in tutte e due le lingue.



Fig(14). Valutazione dell'estrazione di termini equivalenti

Termine italiano	Termine arabo
oltraggio a pubblico ufficiale	tEdy ElY ms&wl EAm
polizia di rapido intervento	$rTp tdxl sryE
ordinanza presidenziale	>mr r}Asy
detenzione preliminare	AHtjAz wqA}y
evasione di massa	frAr jmAEy
crimine di omicidio	jrymp qtl
stupro di massa	AgtSAb jmAEy
rilascio condizionale	<frAj m$rwT
verdetto di colpevolezza	Hkm <dAnp
disobbedienza civile	ESyAn mdny
custodia preprocessuale	Hjz ElY *mp tHqyq
abuso di ufficio	AstglAl mnSb
testimone oculare	$Ahd EyAn
corte di appello	mHkmp Ast}nAf
discriminazione razziale	tmyyz EnSry
giudice co-inquirente	qADy tHqyq m$Ark
aborto forzato	<jhAD qsry
istanza di asilo	Tlb ljw}
giurisdizione militare	qDA' Eskry
separazione sociale	fSl AjtmAEy
giustizia civile	qDA' mdny
cospirazione criminale	t|mr jnA}y
indagine penale	tHqyq jnA}y
sentenza arbitrale	qrAr tHkym
effetto retroattivo	>vr rjEy
indennità di disoccupazione	<EAnp bTAlp
rilascio anticipato	<frAj mbkr
mandato di perquisizione	m*krp tfty$
riduzione di pena	txfyf Eqwbp

Tabella(24). Esempi di termini bilingui estratti

Capitolo VI: Estrazione e analisi delle variazioni terminologiche

6.1. Estrazione delle variazioni terminologiche

L'obiettivo di questa parte del lavoro è individuare le varianti delle terminologie bilingui estratte nella parte precedente. Prima di procedere alla descrizione dell'approccio adottato nella tesi al fine di identificare la varie forme di varianti nel nostro corpus, vediamo opportuno soffermarci brevemente su alcune tecniche di estrazione di varianti da corpora.

Estrarre le diverse varianti delle unità lessicali in un testo rappresenta recentemente uno dei temi cardine nell'ambito del TAL considerata la sua rilevanza per altre applicazioni come l'*information retrieval*, la traduzione automatica, o la creazione delle ontologie, ecc.. Lo stato dell'arte delle tecniche adoperate per trattare le varianti terminologiche dimostra l'adozione di una pluralità di approcci differenti che si basano in senso generale sulla distribuzione contestuale dei termini. L'idea di base di questi approcci è che le parole con significato approssimativamente simile tendono ad avere contesti simili, sia a livello monolingue che plurilingue. La maggior parte degli approcci che sostengono quest'ipotesi si ispira al pensiero di Harris che è stato fra i primi a segnalare la relazione tra le proprietà distribuzionali delle parole in contesto e la loro entità semantica secondo il principio: "the linguistic meanings which the structure carries can only be due to the relations in which the elements of the structure take part"[274].

Le informazioni contestuali di un termine in questo caso si possono fornire o dalle strutture sintattiche delle frasi[275], o dalla gamma delle parole che ricorrono con esso con una certa frequenza[276]. Inoltre, per misurare il contesto di parole si può avvalersi di diversi tipi di risorse lessicali, come per es. corpora, dizionari monolingue o plurilingue, ontologie, ecc..

Di seguito esponiamo alcuni esempi di quelle tecniche:

---

274 Harris, Z., *Mathematical structures of language,* New York, Interscience Publishers, 1968, p.2

275 Cfr. Lin, D.,"Automatic retrieval and clustering of similar words". In *Proceeding COLING '98 Proceedings of the 17th international conference on Computational linguistics - Volume 2*

276 Schütze, H., "Dimensions of meaning". In *Proceedings of the ACM/IEEE conference on Supercomputing,* USA, 1992

Fra i primi lavori rivolti a riconoscere le variazioni dei termini nei corpora tramite mezzi automatici viene in primo luogo il sistema FASTR di Jacquemin[277]. FASTR utilizza una serie di regole logiche (metarules) generate automaticamente da termini annotati e dalle relative strutture morfosintattiche. Le regole, che prendono la forma di schemi liberi dal contesto (context-free portion), rappresentano dei pattern per creare potenziali varianti di termini. Esponiamo di seguito alcuni esempi di queste regole logiche adottate dal sistema per estrarre le tre principali tipologie di variazioni sintattiche dei termini: coordinazione, inserzione, e permutazione:
- regole di coordinazione: ($X_1 \rightarrow X_2\ X_3\ X_4$) = $X_1 \rightarrow X_2\ \mathbf{C_5}\ \mathbf{X_6}\ X_3\ X_4$):
*inflammatory **and erosive** joint disease [inflammatory joint disease]*
- regole di inserzione: ($X_1 \rightarrow X_2\ X_3\ X_4$) = $X_1 \rightarrow X_2\ \mathbf{X_5}\ X_3\ X_4$):
*impaired **intravenous** glucose tolerance [impaired glucose tolerance]*
- regole di permutazione: ($X_1 \rightarrow X_2\ X_3\ X_4$) = $X_1 \rightarrow X_4\ \mathbf{X_5}\ \mathbf{X_6}\ \mathbf{X_7}\ X_2\ X_3$):
*diseases **of the central** nervous system [Nervous system diseases]*

Nelle regole precedenti il lato destro dell'equazione rappresenta la struttura del termine composto originale, mentre il lato sinistro indica l'ordine delle diverse varianti. Il segno → sta per la concatenazione dei costituenti, *X* rappresenta qualsiasi categoria lessicale, e *C* indica qualsiasi congiunzione di coordinazione.
FASTR, malgrado le sue limitazioni (il sistema non riesce a identificare per esempio le varianti semantiche), è stato adoperato per estrarre varianti da corpora francesi e inglesi[278] e da corpora giapponesi[279].

277 Jacquemin, C., Royaute, J., "Retrieving terms and their variants in a lexicalized unification-based framework". In *Proceeding SIGIR '94 Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval,* 1994
278 Cfr Jacquemin, C. "A symbolic and surgical acquisition of terms through variation". In Wermter, S., et al., (a cura di) *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, Heidelberg: Springer, 1996; Jacquemin, C. et al., "Expansion of multi-word terms for indexing and retrieval using morphology and syntax." In Proceedings, 35th Annual Meeting of the Association for Computational Linguistics, Stroudsburg, USA, ACL, 1997
279 Yoshikane, F., et al., "Detecting Japanese Term Variation in Textual Corpus". In *Proceedings 4th International Workshop on Information Retrieval with Asian Languages (IRAL'99*), Taipei, Taiwan, Academia Sinica

Hamon&Nazarenko[280] hanno presentato un approccio finalizzato a individuare le relazioni di sinonimia tra i termini. Si tratta di un metodo basato su regole generative nonché su risorse lessicali esterne (dizionario di lingua generale, tesauro tecnico, classi semantiche costruite manualmente). L'approccio è suddiviso in tre fasi: estrazione di termini da corpora, creazione di nessi semantici tra i termini estratti con l'aiuto delle risorse lessicali, e poi la valutazione delle relazioni semantiche tramite esperti di dominio. Secondo i ricercatori tra due termini esiste una relazione di sinonimia se entrambi i termini in relazione ad un certo contesto sono sintatticamente identici e semanticamente sostituibili. Lavorando solo su termini composti, il sistema suddivide ogni termine in "testa" e "espansione", e considera sinonimi due termini se i loro relativi costituenti sono identici o sinonimi. Quindi dati due termini composti $\mathbf{CCT}_1(\mathbf{T}_1,\mathbf{E}_1)$ e $\mathbf{CCT}_2(\mathbf{T}_2,\mathbf{E}_2)$ e la relazione di sinonimia syn ($\mathbf{CT}_1$, $\mathbf{CT}_2$) tra i due termini candidati $\mathbf{CT}_1$ e $\mathbf{CT}_2$, si applicano le seguenti tre regole di sinonimia:

R_1: $\mathbf{T}_1 = \mathbf{T}_2 \wedge \text{syn}(\mathbf{E}_1,\mathbf{E}_2) \supset \text{syn}(\mathbf{CCT}_1, \mathbf{CCT}_2)$

R_2: $\mathbf{E}_1 = \mathbf{E}_2 \wedge \text{syn}(\mathbf{T}_1,\mathbf{T}_2) \supset \text{syn}(\mathbf{CCT}_1, \mathbf{CCT}_2)$

R_3: $\text{syn}(\mathbf{T}_1, \mathbf{T}_2) \wedge \text{syn}(\mathbf{E}_1,\mathbf{E}_2) \supset \text{syn}(\mathbf{CCT}_1, \mathbf{CCT}_2)$

Questo vuol dire che tra due termini candidati si può avere una relazione di sinonimia se viene soddisfatta una delle seguenti condizioni:

- le teste sono identiche e le espansioni sono sinonime, come *collecteur général* / *collecteur commun*;
- le teste sono sinonime e le espansioni sono identiche *matériel électrique / équipement électrique;*
- le teste sono sinonime e le espansioni sono sinonime *marche normale / bon functionnement*

Il sistema utilizza preliminarmente le relazioni semantiche fornite dal dizionario come mezzo per identificare casi di sinonimia tra i termini, che si utilizzeranno iterativamente poi per determinare nuove relazioni di sinonimia.

280 Hamon T., Nazarenko, A. "Detection of synonymy link between terms: Experiment and results". In Bourigault D., et al (a cura di), *Recent Advances in Computational Terminology*, volume 2 of Natural Language Processing John Benjamins, 2001

Nell'ambito della piattaforma T2K[281] si è proceduto a realizzare una strutturazione concettuale del glossario terminologico creato in una fase iniziale. Le principali relazioni semantiche individuate in questo lavoro sono:
a) iponimia;
b) affinità semantica

L'identificazione della relazione di iponimia tra i termini composti si basa essenzialmente sul concetto di inclusione lessicale:
Due unità terminologiche polirematiche che condividano la medesima testa lessicale (e talora anche gli stessi modificatori) al livello della loro rappresentazione in chunks vengono interpretati come iponimi del termine corrispondente alla struttura condivisa[282].
Per riconoscere invece la relazione di affinità semantica si è adottato un metodo basato effettivamente sulle proprietà distribuzionali dei termini nei corpora. Secondo questo approccio "due parole sono semanticamente simili se sono reciprocamente sostituibili in un numero significativo di contesti sintattici", come per es.
-*abrogare un decreto* e *abrogare una direttiva*
-*integrare un decreto* e *integrare una direttiva*

Quindi il fatto che *decreto* e *direttiva* co-occorrono con la stessa funzione sintattica con gli stessi verbi (*abrogare* e *integrare*) sta per indicare che le due parole sono semanticamente simili.

Fra gli approcci che utilizzano risorse plurilingui è quello proposto da Bannard e Callison-Burch[283]. Gli autori cercano di estrarre parafrasi da corpora monolingui tramite il confronto delle loro relative traduzioni in corpora paralleli bilingui o plurilingui. Il nucleo dell'approccio consiste nell'identificare le diverse varianti dei termini in una lingua originale che hanno un unico corrispondente nella lingua d'arrivo, come dimostra l'esempio seguente:

281 Dell'Orletta, F. et al., "Dal testo alla conoscenza e ritorno" op cit.
282 Ivi, p.198
283 Bannard, C., Callison-Burch, C., "Paraphrasing with bilingual parallel corpora". In *Proceedings of the annual Meeting of the Association for Computational Linguistics (ACL)*, Ann Arbor, MI, 2005

What is more, the relevant cost dynamic is completly **under control**
Im übrigen ist die diesbezügliche kostenentwicklung völlig **unter kontrolle**
Wir sind es den steuerzahlern schuldig die kosten **unter kontrolle** zu haben
We owe it to the taxpayers to keep the costs **in check**

Nell'esempio sopracitato le due espressioni inglesi *under control* e *in check* rappresentano due parafrasi perché hanno in comune un unico corrispettivo in tedesco *unter kontrolle.*

Dopo l'allineamento dei corpora paralleli a livello di parole, si cerca di selezionare coppie di parafrasi tramite l'assegnamento di probabilità di parafrasi in questo modo:

$\hat{e}_2 = \arg\max_{e_2 \neq e_1} p(e_2|e_1)$
$= \arg\max_{e_2 \neq e_1} \sum_f p(f|e_1)\, p(e_2|f)$, dove

e_1 è l'espressione nella lingua d'origine, e_2 è la parafrasi candidata nella lingua d'origine, f è la traduzione dell'espressione nella lingua d'arrivo.
Le parafrasi selezionate con maggiori probabilità vengono poi filtrate tramite un modello di linguaggio che utilizza come informazione contestuale la frase in cui occorre e_1.

In un altro lavoro[284] Callison-Burch ha esteso il filtro delle parafrasi applicando un ulteriore criterio sintattico consistente nell'accettare solo le parafrasi che concordino sintatticamente con le espressioni originali.

In un simile lavoro, Van Der Plas e Tiedemann[285] hanno dimostrato che ci si può servire anche delle informazioni fornite dal PoS tagging per filtrare le parafrasi candidate.

284 Callison-Burch, C.,"Syntactic Constraints on Paraphrases Extracted from Parallel Corpora". In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii. Association for Computational Linguistics.
285 Van Der Plas L. , Tiedemann J., "Finding medical term variations using parallel corpora and distributional similarity". In: *Proceedings of the Coling workshop on ontologies and lexical resources,* Beijing, Cina, 2010.

Per ottimizzare la precisione e la *recall* dei metodi che utilizzano un solo tipo di risorse lessicali, Wu&Zhou[286] hanno proposto un approccio complementare all'estrazione di relazioni di sinonimia combinando tre risorse lessicali diverse: dizionario monolingue, dizionario bilingue e corpus bilingue. Inizialmente si cerca di creare relazioni di sinonimia in modo separato da ciascuna risorsa lessicale:

a) nel caso del dizionario monolingue si adoperano le parole che compongono la definizione di ogni singola voce (*hubs*) nonché la partecipazione della voce stessa ad altre definizioni di altre voci (*authorities*) come vettori di caratteristica (*feature vector*) per rappresentare semanticamente un termine, partendo dall'assunto che due termini sono semanticamente simili se hanno in comune *hubs* e *authorities;*

b) nel caso del corpus parallelo si utilizzano le traduzioni dei termini per esprimere il loro significato basandosi sull'idea che due termini sono sinonimi se le loro relative traduzioni appaiono simili. In questa fase ci si serve preliminarmente di un dizionario bilingue per rapportare ogni voce nella SL alle sue traduzioni nella TL. Successivamente a ciascuna di quelle traduzioni viene assegnata una probabilità acquisita da un corpus parallelo allineato a livello di parole.

c) la terza fase comprende l'utilizzo di un corpus monolingue per individuare relazioni di sinonimia tra i termini in base al loro contesto che in questo caso è l'insieme delle parole con cui ciascun termine ha una relazione di dipendenza. Per avere quest'ultima informazione bisogna analizzare il corpus a livello sintattico nel formato triplo seguente: <parola1, Tipo Relazione, parola2 >. Gli ultimi due elementi del triplo rappresentano gli attributi del primo elemento. In questo caso due parole sono sinonime se condividono gli stessi attributi.

286 Wu, H., Ming Zhou, M.,"Optimizing synonym extraction using monolingual and bilingual resources. In *Proceedings of the second international workshop on Paraphrasing,* Association for Computational LinguisticsStroudsburg, PA, USA, 2003

In una fase finale l'output dei tre metodi precedenti viene unificato tramite un metodo di apprendimento d'insieme (*ensemble learning method*) per misurare statisticamente la relazione di sinonimia tra coppie di parole.

Riflettendo sulla rassegna panoramica precedente possiamo sottolineare due osservazioni: a) la maggior parte degli approcci si interessa di identificare relazioni semantiche, soprattutto di sinonimia, tra i termini; b) le variazioni terminologiche vengono affrontate solo da un punto di vista intralinguistico, il che significa che le variazioni si estraggono solo da testi monolingui anche se vengono coinvolti altri corpora plurilingui per migliorare i risultati.

6.2. L'approccio della tesi all'estrazione delle variazioni terminologiche

Il processo di estrazione delle varianti si configura come un metodo semiautomatico. In una prima fase, viene adoperato un sistema computazionale che utilizza le informazioni dei termini bilingui selezionati e dei termini candidati estratti nella fase iniziale del lavoro. In una fase finale si adotta un intervento manuale finalizzato a filtrare i risultati del codice informatico.

Il codice adottato per estrarre le varianti dei termini selezionati passa per le seguenti fasi:

1- In una prima fase si utilizza il lemma dei termini selezionati e dei termini candidati: se due termini condividono una testa e/o un modificatore e risultano sostituibili a livello del corpus parallelo (la sostituibilità viene misurata in questo caso tramite la presenza di un unico corrispettivo nell'altra lingua), vengono considerati varianti, come dimostrano gli esempi seguenti:

Termine italiano	Forma lemmatizzata	Varianti	Termine arabo	Forma lemmatizzata	Varianti
disposizioni di legge	disposizione di legge	disposizioni legali	>HkAm AlqAnwn	Hkm qAnwn	>HkAm qAnwnyp
		disposizione contenuta nella legge			>HkAm AlqAnwn AljnA}y
		violazione delle disposizioni di legge			Al>HkAm AlwArdp fy AlqAnwn
diritti umani	diritto umano	diritto umanitario	Hqwq Al<nsAn	Hq <nsAn	Hqwq Al>$xAS
		diritti civili e umani			Hqwq Al<nsAn AlsAsyp
		diritti fondamentali della persona			Hqwq kl <nsAn
		diritti dell'uomo			
diritti di proprietà	diritto di proprietà	diritto alla proprietà	Hq Almlkyp	Hq mlkyp	AlHq fy Almlkyp

2- Per le varianti, soprattutto quelle semantiche, che il passo precedente non riesce a individuare, si usa qui un codice che utilizza le seguenti informazioni:

- i termini bilingui selezionati;
- il corpus parallelo;
- i termini candidati;
- WordNet[287] che è una banca dati semantico-lessicali per la lingua inglese;
- la traduzione in inglese dei termini bilingui nonché dei termini candidati, tramite un sistema SMT.

287 WordNet è disponibile gratuitamente per la comunità scientifica al sito dell'università di Princeton: https://wordnet.princeton.edu

L'idea base di questo metodo parte dell'ipotesi che tramite l'utilizzo di un sistema concettuale- semantico come WordNet in una lingua pivot, che è in questo caso l'inglese, si possano identificare le varianti semantiche in due lingue diverse. WordNet, considerato una delle risorse lessicali più importanti per la lingua inglese, raggruppa e organizza sostantivi, verbi, aggettivi e avverbi in insiemi di sinonimi cognitivi, detti *synset*, ognuno dei quali esprime un concetto distinto. I synset sono strutturati in nodi in base alle relazioni concettuale-semantiche diverse: iperonimia, iponimia, mereonimia, sinonimia, ecc.. Nel nostro lavoro ci interessa maggiormente la relazione di sinonimia. Per quanto riguarda l'adozione di un sistema SMT per trovare i corrispettivi in inglese dei termini possiamo dire che rispetto alla traduzione testuale in cui si possono riscontare errori dovuti principalmente all'ambiguità semantica, nel caso di singole unità terminologiche la traduzione prestata presenta un alto livello di precisione.
L'esempio seguente spiega il procedimento di questo passo dell'approccio:

- prendiamo i due termini bilingui della fase precedente: "disposizioni della legge", ">HkAm AlqAnwn"

- utilizziamo Google Translate per tradurre entrambi i termini in inglese. In questo caso otteniamo "provisions of the law" come traduzione per ambedue i termini italiano-arabi. Bisogna dire che in questo caso i due termini bilingui possono avere come corrispondente in inglese due forme lessicali diverse;

- si cercano, poi, in WordNet database i termini che abbiano una relazione concettuale-semantica con il termine ottenuto tramite il SMT, senza considerare ovviamente le parole funzionali, il che significa che nel caso dell'esempio in questione si cercano i synset solo per le due parole "provisions" e "law". In questa maniera si hanno questi termini connessi semanticamente (synset):
{'planning', 'natural_law', 'practice_of_law', 'viands', 'jurisprudence', 'police_force', 'provision', 'purvey', 'law', 'constabulary', 'commissariat', 'supply', 'preparation', 'law_of_nature', 'provisions', 'victuals', 'legal_philosophy', 'proviso', 'police', 'provender', 'supplying'}

- nella stessa maniera si crea per ogni termine candidato un dizionario che contiene i relativi synset. L'idea di applicare questa fase ai termini candidati estratti inizialmente tramite i pattern linguistici prima del filtro statistico è basata sulla possibilità di garantire una certa *recall* dei risultati.

- si cercano poi a livello monolingue corrispondenze semantiche tra i termini selezionati e i termini candidati tramite i relativi synset creati nelle due fasi precedenti. In questo senso l'uguaglianza concettuale si realizza quando si verifica la compresenza di uno o più di elementi nei due synset.

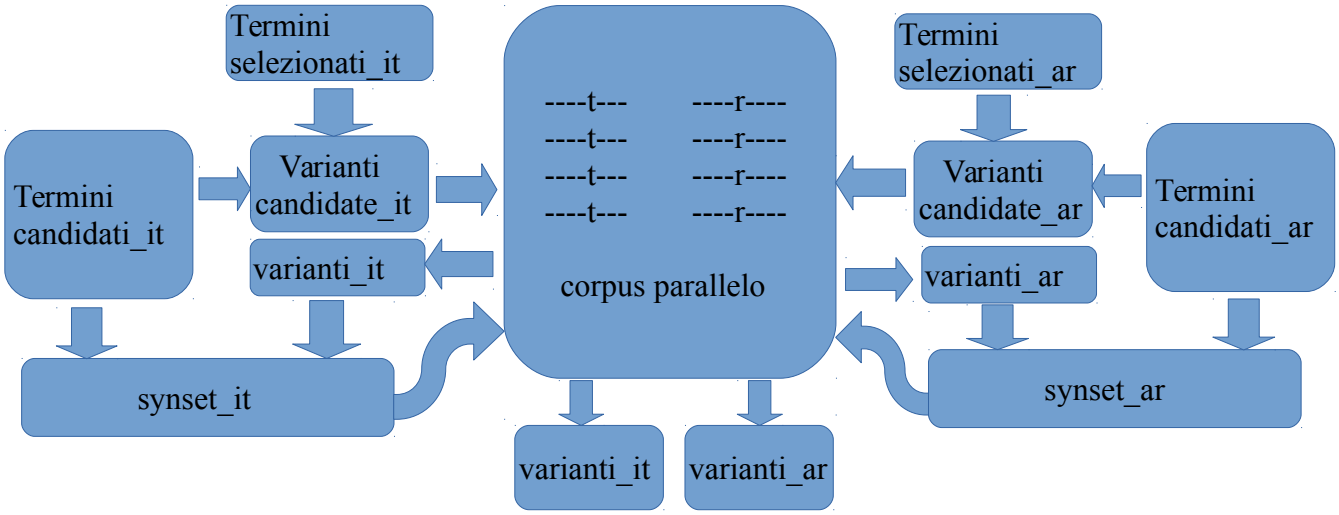
- si utilizza in una fase finale il corpus parallelo per verificare la sostituibilità tra i termini base e le relative varianti.

Quindi dopo l'applicazione di questa tappa risultano di una certa relazione semantica con il termine italiano "disposizioni della legge" le seguenti forme lessicali selezionati dai termini candidati:

- disposizioni applicabili della legislazione
- disposizioni legislative
- prescrizioni della legislazione nazionale
- prescrizioni della legislazione
- provvedimenti legislativi
- sensi di legge

mentre per il termine arabo ">HkAm AlqAnwn" si ottengono le seguenti varianti semantiche:

- mwAd AlqAnwn
- >HkAm Alt$ryE
- AltdAbyr Alt$ryEyp
- AlmwAd AlqAnwnyp
- bnwd AlqAnwn
- AlnSwS Alt$ryEyp



Fig(15). Architettura del metodo per estrarre varianti

if termine_selezionato e termine_candidato condividono una testa **and/or** un modificatore **and** risultano sostituibili a livello del corpus parallelo:
    varianti.add(termine_candidato)

for variante in varianti:
    for elemento in termini_candidati
        for synset in variante_synset:
            for synset_ in elemento_synset:
                if synset e synset_ condividono un item **and** variante e elemento risultano sostituibili a livello del corpus parallelo:
                    varianti.add(item)

Fig(16): l'algoritmo dell'estrazione delle variazioni terminologiche

6.3. Risultati delle variazioni

Dopo l'esclusione dei termini che non hanno presentato varianti, i risultati ottenuti sono classificati nella maniera seguente: dalle varianti estratte si seleziona la forma più frequente per rappresentare il termine base. Le diverse variazioni verranno etichettate poi a seconda della relazione stabilita con il termine base. Le tipologie principali considerate nel nostro studio sono quattro: semantiche, morfo-sintattiche, sintattiche e ortografiche.

Le variazioni semantiche consistono principalmente in due classi: sinonimia parziale e sinonimia. Mentre la sinonimia parziale riguarda la sostituzione di una o più componenti essenziali (escluse cioè le parole funzionali) del termine base con altre forme lessicali, la sinonimia si realizza quando si sostituiscono tutti i costituenti del termine base. Per es. rispetto al termine base *rimpatri forzati*, la variante *espulsioni non spontaneo* rappresenta una relazione di sinonimia, bensì *espulsioni forzate* oppure *rimpatri non spontanei* sono sinonimia parziale.

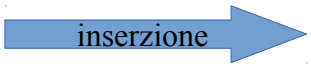
Le variazioni morfo-sintattiche nel nostro lavoro concernono essenzialmente la morfologia derivazionale dei termini base, trattandosi cioè del cambiamento della categoria morfologica dei termini, come per es.

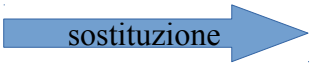
rimpatri forzati — da aggettivo a nome → *rimpatri con la forza;*

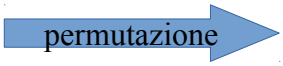
decreto del presidente — da nome a aggettivo → *decreto presidenziale*

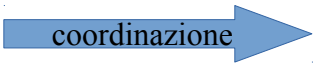
Le variazioni del tipo sintattiche riguardano, invece, direttamente la struttura interna dei costituenti dei termini composti, implicando cioè cambiamento all'interno di questa struttura. Il cambiamento in questo caso potrebbe essere provocato da inserzione di altri elementi, da sostituzione di qualche parola funzionale, da permutazione dell'ordine degli elementi stessi, dall'omissione di una componente, o dalla coordinazione di qualche elemento del termine base a altri elementi lessicali. L'esempio seguente indica le

tipologie delle variazioni sintattiche considerate nella tesi:

violenza domestica  violenze **in ambito** domestico;

processo di corruzione  processo **per** corruzione;

polizia di rapido intervento  polizia di **intervento rapido**;

violenza domestica  violenza **sessuale** e domestica

consiglio supremo della magistratura  consiglio della magistratura

La tabella seguente dimostra i primi 5 termini bilingui con le loro relative variazioni:

	Termine di base italiano	Varianti del Termine di base italiano	Tipo varianti	Termine di base arabo	Varianti del Termine di base arabo	Tipo varianti
1	pena di morte	**braccio** della morte	sinonimia parziale	Hkm Al<EdAm	Hkm **b**Al<EdAm (condanna a morte)	inserzione
		condanna a morte	sinonimia parziale		**EmlyAt** Al<EdAm (azioni di esecuzione)	sinonimia parziale
		pena **capitale**	sinonimia parziale			
		sentenza di morte	sinonimia parziale			
2	violenza domestica	violenza **in ambito familiare**	inserzione+sinonimia parziale	AlEnf Al>sry	AlEnf **Dd Almr>p fy mHyT Al>srp** (violenza contro donna all'interno della famiglia)	Inserzione + morfo-sintattico
		violenza **in ambito** domestico	Inserzione		AlEnf **fy Al<TAr** Al>sry (violenza nell'ambito familiare)	inserzione
		violenza **all'interno della famiglia**	inserzione+sinonimia parziale		AlEnf **fy <TAr Al>srp** (violenza in ambito della famiglia)	Inserzione + morfo-sintattico
		violenza **familiare**	sinonimia parziale		AlEnf **fy mHyT Al>srp** (violenza in ambito della famiglia)	Inserzione + morfo-sintattico
					AlEnf **Almnzly** (violenza domestica)	sinonimia parziale

		violenza **sessuale e** domestica	coordinazione		**>EmAl** AlEnf **AljnA}yp fy mHyT Al>srp** (atti di violenza all'interno della famiglia)	Inserzione + morfo-sintattico
					AlEnf **dAxl Al>srp** (violenza dentro la famiglia)	Inserzione + morfo-sintattico
3	grazia presidenziale	**Amnistia** presidenziale	sinonimia parziale	Efw r}Asy	Efw **mn Alr}ys**	Inserzione + morfo-sintattico
		grazia **del presidente**	morfosintattico			
		Clemenza da parte del presidente	sinonimia parziale + inserzione +morfosintattico			
4	funzionario pubblico	funzionari **statali**	sinonimia parziale	mwZf Emwmy	mwZfy **AlHkwmp** (impiegati del governo)	sinonimia parziale
		impiegati pubblici	sinonimia parziale		**mstxdmw Aldwlp** (dipendenti del paese)	sinonimia
		funzionari **governativi**	sinonimia parziale		mwZfy **Aldwlp** (impiegati del paese)	sinonimia parziale
		pubblici **ufficiali**	sinonimia parziale		mwZfyn **fy Aldwlp** (impiegati nel paese)	sinonimia parziale
		funzionari **delle amministrazioni** pubbliche	inserzione			
		agente statale	sinonimia			
5	minoranze etniche	**gruppi** etnici	sinonimia parziale	>qlyAt Erqyp	Al>qlyAt **Al<vnyp** (minoranza etnica)	sinonimia parziale
		minoranza **razziale**	sinonimia parziale		Al>qlyAt **AlEnSryp w**AlErqyp (minoranza razziale e etnica)	coordinazione
		minoranze **razziali ed** etniche	coordinazione		**AljmAEAt** AlErqyp (gruppi etnici)	sinonimia parziale
		gruppi razziali ed etnici	sinonimia parziale + coordinazione		**AlmjtmEAt Al<vnyp** (comunità etniche)	sinonimia
		comunità etniche	sinonimia parziale		Al>qlyAt **AlAvnyp w** AlErqyp (minoranza razziale e etnica)	coordinazione

Tabella (25): Alcuni esempi delle variazioni terminologiche estratti dalla lista dei risultati

6.3.1 Variazioni semantiche

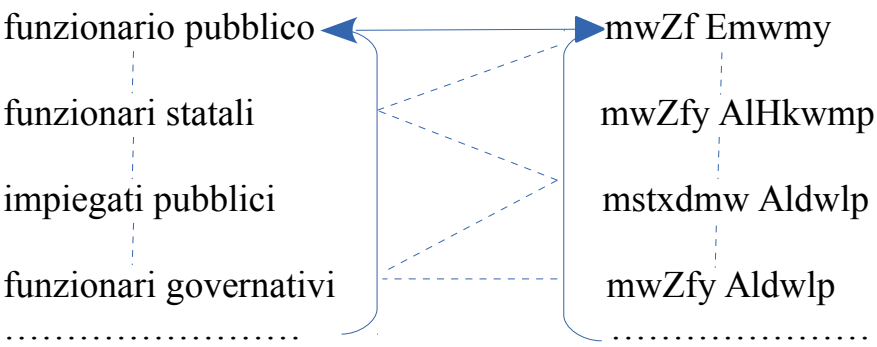
Come è stato anticipato le variazioni semantiche nel nostro lavoro si limitano alla relazione di sinonimia, sia parziale che completa.

Mentre la relazione di sinonimia parziale viene realizzata quando una delle componenti (testa o modificatore[288]) del termine composto viene sostituita con un'altra variante lessicale, la sinonimia avrà luogo nel caso della sostituzione di entrambe le componenti del termine.

Come è stato descritto nella parte riguardante l'approccio adottato per individuare le varianti, il criterio base per riconoscere una forma lessicale come variante semantica di un termine base nella SL è la sua presenza, con una certa frequenza, come corrispondente di un termine, o una sua variante, nella TL, che a sua volta deve essere, in un modo diretto o indiretto, corrispondente del termine base nella SL. Questo spiega il ruolo fondamentale del corpus parallelo nel riconoscere in una maniera incrementale le diverse varianti dei termini. L'esempio seguente dimostra l'andamento di questo approccio:
funzionario pubblico ↔ *mwZf Emwmy* sono due termini bilingui estratti come corrispondenti. Nelle frasi allineate nel corpus parallelo si guarda se come corrispondente del termine *mwZf Emwmy* è un altro termine *x*, presente nella lista dei termini candidati italiani e diverso da *funzionario pubblico.* In quel caso si considera *x* come variante semantica di *funzionario pubblico.* Poi in una maniera simile si guarda nelle frasi bilingui per sapere se come equivalenti di *funzionario pubblico* e *x* c'è un'altra forma lessicale *y*, nella lista dei termini candidati arabi, diversa da *mwZf Emwmy* e in questo caso si ritiene *y* come variante semantica, o sinonimia, dell'unità lessicale *mwZf Emwmy,* e così via finché non si incontrano nuovi forme lessicali nel corpus parallelo.

---

288 Nell'ambito dei costrutti sintagmatici la testa è la parte responsabile della determinazione semantica e grammaticale di tutto il costrutto, mentre il modificatore ha la funzione di restringere l'estensione della testa aggiungendo ulteriori aspetti al suo significato. Cfr. Grossmann, M.&Rainer, F., *La formazione delle parole in italiano*, op cit. pp 11:12



corpus_parallelo.xml

italiano	arabo
Ogni Stato Parte si impegna a proibire in ogni territorio sotto la sua giurisdizione altri atti costitutivi di pene o trattamenti crudeli, inumani o degradanti che non siano atti di tortura quale definita all articolo 1, qualora siano compiuti da un **funzionario pubblico** o da qualsiasi altra persona che agisce a titolo ufficiale, o sotto sua istigazione, oppure con il suo consenso espresso a tacito.	تتعهد كل دولة طرف بأن تمنع، في أي إقليم يخضع لولايتها القضائية حدوث أي أعمال أخرى من أعمال المعاملة أو العقوبة القاسية أو اللاإنسانية أو المهينة التي لا تصل إلى حد التعذيب كما حددته المادة 1، عندما يرتكب **موظف عمومي** أو شخص آخر يتصرف بصفة رسمية هذه الأعمال أو يحرض على ارتكابها، أو عندما تتم بموافقته أو بسكوته عليها.
Ogni Stato Parte prende in considerazione l adozione di misure legislative e di altre necessarie a conferire il carattere di reato agli atti di cui al paragrafo 1 del presente articolo, che coinvolgono un **pubblico ufficiale** straniero o un funzionario internazionale.	تنظر كل دولة طرف في اعتماد ما قد يلزم من تدابير تشريعية وتدابير أخرى لتجريم السلوك المشار إليه في الفقرة 1 من هذه المادة الذي يكون ضالعا فيه **موظف عمومي** أجنبي أو موظف مدني دولي.
Tutti e sette sono stati accusati di "teppismo" e aggressione a **pubblico ufficiale**, ma sei sono stati in seguito rilasciati su cauzione..	وقد وُجهت إلى الأشخاص السبعة جميعاً تهمة الشغب والاعتداء على **موظفي الدولة**، ولكن تم إطلاق سراح ستة منهم بكفالة في وقت لاحق.
Dovevano rispondere di accuse come violenza o intimidazione nei confronti di un **funzionario statale**, che ha in seguito ritirato l'accusa.	وقد واجه الزوجان تهماً باستخدام العنف أو الترهيب ضد أحد **موظفي الدولة**، ولكنه سحب الاتهام في وقت لاحق.
Ad agosto, ha destato scalpore nell opinione pubblica la proposta di un **funzionario statale** affinché l Università Teknologi Mara (UITM) stanziasse il 10% dei posti universitari a non malay.	ففي أغسطس/آب، ثار غضب شعبي عارم بعد أن اقترح أحد **المسؤولين الحكوميين** أن تخصص «جامعة مارا التقنية» 10 بالمئة من الأماكن بالجامعة لمن ينتمون لجماعات أخرى غير «الماليا».
Funzionari governativi hanno ammesso che ciò costituiva un grave problema e hanno chiesto pene più severe per coloro che venivano riconosciuti colpevoli di questo tipo di crimini.	وأقر بعض **المسؤولين الحكوميين** بأن ذلك يمثل مشكلة خطيرة، ودعوا إلى فرض عقوبات قاسية على من يُدانون بارتكاب مثل هذه الجرائم.
I sospettati di un possibile coinvolgimento di **funzionari governativi** nella sua scomparsa hanno provocato proteste nella stampa, tra la società civile e tra i partiti politici di opposizione.	وقد أثارت الشكوك المتعلقة باحتمال ضلوع **موظفي الحكومة** في حادثة اختفائه احتجاجات من قبل الصحافة ومنظمات المجتمع المدني والأحزاب السياسية المعارضة.
A novembre, quattro membri del Sindacato di categoria dei dipendenti pubblici camerunesi, Centrale Syndicale du Secteur Public, tra cui il presidente dello stesso, Jean Marc Bikoko, e la vice presidente, Brigitte Tamo, sono stati arrestati da gendarmi durante una manifestazione pacifica in cui veniva chiesto l aumento dello stipendio dei **dipendenti pubblici**.	في نوفمبر/تشرين الثاني، اعتقل أفراد الدرك أربعة أعضاء في النقابة العامة **للعاملين في القطاع العام**، ومن بينهم رئيس النقابة جان مارك بيكوكو ونائبة الرئيس بريجيت تامو، خلال مظاهرة سلمية للمطالبة بزيادة أجور موظفي الحكومة.
La pressione esercitata sui **lavoratori dipendenti pubblici**, insegnanti compresi, ha portato a lunghi scioperi.	وأدت الضغوط التي مورست على **العاملين في القطاع العام**، بمن فيهم المعلمون، إلى تنفيذ إضرابات طويلة.
Diverse persone sono rimaste ferite e una persona è rimasta uccisa nel corso di manifestazioni indette da **lavoratori del settore pubblico** per chiedere l aumento dei salari e miglioramenti delle condizioni di lavoro.	جُرح عدة أشخاص وقُتل شخص واحد خلال مظاهرات **لعمال القطاع العام** للمطالبة برفع الأجور وتحسين الأوضاع.

Fig.(17) estratto del corpus parallelo che dimostra le varianti di sinonimia dei due termini bilingui *funzionario pubblico* ↔ *mwZf Emwmy*

Le tabelle seguenti dimostrano i risultati della sinonimia. Nella prima colonna si riporta il termine base italiano, nella seconda colonna ci sono le varianti del termine base italiano, nella terza colonna sta il numero di queste varianti. Le altre tre colonne della lingua araba seguono lo stesso ordine dell'italiano. In quanto sono corrispondenti dei termini base italiani presenti nella stessa riga, i termini arabi e le loro varianti non vengono accompagnati da una traduzione in italiano.

6.3.1.1. Sinonimia parziale con sostituzione di testa

Termine base italiano	Varianti del termine base italiano	n	Termine di base arabo	Varianti del termine base arabo	n
pena di morte	braccio della morte/condanna a morte/sentenza di morte	3	Hkm Al<EdAm	EmlyAt Al<EdAm/Eqwbp Al<EdAm	2
grazia presidenziale	amnistia presidenziale	1	Efw r}Asy	--	
decreto presidenziale	ordinanza presidenziale	1	mrswm r}Asy	qrAr r}Asy/>mr r}Asy	2
funzionario pubblico	impiegato pubblico/dipendente pubblico/pubblico ufficiale	3	mwZf Emwmy	ms&wlyn Emwmyyn	1
minoranze etniche	gruppi etnici/comunità etniche	2	>qlyAt Erqyp	jmAEAt Erqyp	1
detenzione arbitraria	arresto arbitrario	1	AlAEqAl AlEsfy	AlqbD AlEsfy/AlAHtjAz AlEsfy	2
rimpatri forzati	espulsioni forzate/trasferimento forzato	2	AlAbEAd Alqsry	AlArjAE Alqsry/ AlAEAdp Alqsryp/ AlrHyl Alqsry	3
sgombero forzato	escomi forzati/sfollamento forzato/sfratti forzati	3	<xlA' qsry	<jlA' qsry/thjyr qsry/trHyl qsry	3
autorità competente	organismo competente/organo competente/amministrazione competente	3	slTp mxtSp	jhp mxtSp/hy}At mxtSp	2
prigioniero di guerra	--		>srY Hrb	sjnA' Hrb	1
procuratore generale	--		AlmdEy AlEAm	/AlnA}b AlEAm/AlAdEA' AlEAm/AlmHAmy AlEAm	3
progetto di legge	disegno di legge/bozza di legge/proposta di legge	3	m$rwE qAnwn	mswdp qAnwn	1
agente di polizia	ufficiale di polizia/funzionario di polizia/forze di polizia	3	DAbT $rTp	frd $rTp/ms&wlw Al$rTp/rjAl Al$rTp	3
violenza sessuale	abuso sessuale/aggressione sessuale/reato sessuale	3	AlEnf Aljnsy	Al<y*A' Aljnsy/AlAEtdA' Aljnsy/AlAnthAk Aljnsy	3
legislazione antiterrorismo	decreto antiterrorismo/normativa antiterrorismo	2	qAnwn mkAfHp Al<rhAb	t$ryE mkAfHp Al<rhAb	1
procedimento giudiziario	atti giudiziari/procedure giudiziarie	2	Al<jrA'At AlqDA}yp	AltdAbyr AlqDA}yp	1
codice penale	legge penale/diritto penale	2	AlqAnwn AljnA}y	AlqDA' AljnA}y/nZAm AlEdAlp	2

				AljnA}yp	
servizio militare	leva militare	1	Alxdmp AlEskryp	Altjnyd AlEskry	1
discriminazione razziale	odio razziale/intolleranza razziale	2	Altmyyz AlEnSry	AlkrAhyp AlEnSryp / AltESb AlEnSry	2
mandato di arresto	procedura di arresto	1	>mr bAlqbD	<*nAF bAlqbD/m*krp qbD	1
richiesta di estradizione	domanda di estradizione	1	Tlb Altslym	--	
attuazione della legge	applicazione della legge	1	tnfy* AlqAnwn	tTbyq AlqAnwn	1
disposizioni di legge	prescrizioni di legge	1	AHkAm AlqAnwn	bnwd AlqAnwn	1
polizia antisommossa	--		$rTp mkAfHp Al$gb	qwAt mkAfHp Al$gb	1
contrattazione collettiva	negoziazione collettiva/trattativa collettiva	2	AlmfAwDp AljmAEyp	AlmsAwmp AljmAEyp/AlAtfAq AljmAEy/AltfAwD AljmAEy	3
detenzione amministrativa	arresto amministrativo	1	AlAEtqAl Al<dAry	AlHbs Al<dAry/AlAHtjAz Al<dAry/>HkAm sjn <dAryp	3
status di rifugiato	stato di rifugiato/condizioni di rifugiato	2	Sfp AllAj}	wDE AllAj}	1
datore di lavoro	--		>SHAb Al>EmAl	>rbAb Al>EmAl	1
legislazione nazionale	legge nazionale/ordinamento nazionale	2	qAnwn wTny	t$ryE wTny/AllwA}H AlwTnyp	2
matrimonio omosessuale	relazione omosessuale/unione civile omosessuale	2	zwAj mvly	ElAqAt jnsyp mvlyp	1
dissidente politico	oppositore politico	1	mEArD syAsy	AlxSwm AlsyAsyyn/mn$q syAsy	2
sistema penitenziario	regime penitenziario	1	nZAm Alsjwn	-	
emendamento alla legge	riforma alla legge /modifica alla legge	2	tEdyl qAnwn	mrAjEp qAnwn/<SlAH qAnwn	2
corte civile	tribunale civile	1	mHkmp mdnyp	--	
risarcimento economico	indennizzo economico	1	AltEwyD AlmAlY	-	
sindacato degli avvocati	ordine degli avvocati	1	nqAbp AlmHAmyyn	AtHAd AlmHAmyyn	1
corte amministrativa	tribunale amministrativa	1	AlmHkmp Al<dAryp	-	
competenza della corte	giurisdizione della corte	1	AxtSAS AlmHkmp	AlwlAyp AlqDA}yp lmHAkm	1
traffico di droga	commercio di droga/reati di droga/spaccio di droga	3	tjArp AlmxdrAt	thryb AlmxdrAt	1
ammissione di colpevolezza	dichiarazione di colpevolezza	1	AlAEtrAf bAl*nb	Al<qrAr bAl*nb	1
giudizio di colpevolezza	verdetto di colpevolezza	1	Hkm Al<dAnp	qrAr Al<dAnp	1
inchiesta ufficiale	indagine ufficiale	1	tHqyq rsmy	-	
organizzazioni della società civile	associazioni della società civile/gruppi della società civile	2	mnZmAt AlmjtmE Almdny	jmAEAt mn AlmjtmE Almdny/hy}At AlmjtmE Almdny	2
codice di condotta	linee di condotta	1	qwAEd Alslwk	mbAd} Alslwk	1
sentenza della corte	decisione della corte/giudizio della corte/verdetto della corte	3	Hkm mHkmp	qrAr AlmHkmp	1
bande criminali	organizzazioni criminali/associazione criminale/gruppi criminali	3	AlESAbAt Al<jrAmyp	AlmjmwEAt Al<jrAmyp/AljmAEAt Al<jrAmyp	2

modifiche costituzionali	mutamenti costituzionali/riforme costituzionali/cambiamenti costituzionali/emendamenti costituzionali	4	AltEdylAt Aldstwryp	Al<SlAHAt Aldstwryp	1
polizia di rapido intervento	unità di reazione rapida	1	$rTp Altdxl AlsryE	ktybp Altdxl AlsryE/qwp Altdxl AlsryE/mjmwEp Altdxl AlsryE	3
oltraggio a pubblica autorità	--		AltEdy ElY ms&wl EAm	Alq*f fy Hq ms&wl EAm	1
pene detentive	sanzioni detentive	1	>HkAm Alsjn	Eqwbp Alsjn	1
diritto internazionale	legislazione internazionale/norme internazionali/giuridiche internazionali/leggi internazionali/giurisprudenza internazionale	5	AlqAnwn Aldwly	AlnwAmys Aldwlyp/AlSkwk Aldwlyp	2
processo di corruzione	caso giudiziario di corruzione	1	qDyp fsAd	jrA}m fsAd	1
tentato omicidio	tentato assassinio	1	Al$rwE fy Alqtl	mHAwlp Alqtl	1
mancanza di prove	insufficienza di prove/assenza di prove	2	Edm kfAyp Al>dlp	Edm twfr Al>dlp	1
fabbricazione di prove	falsificazione di prove	1	tlfyq Al>dlp	tzwyr Al>dlp/ds >dlp	2
occultamento di prove	trattenimento di prove	1	Hjb Al>dlp	<xfA' Al>dlp	1
decesso in custodia	morte in custodia	1	wfAp fy AlHjz	-	
stato d'emergenza	--		HAlp AlTwAr}	>HkAm AlTwAr}	1
doppia cittadinanza	doppia nazionalità	1	jnsyp mzdwjp	jwAzAt sfr mzdwjp	1
richiesta di asilo	istanza di asilo/domanda di asilo	2	Tlb ljw}	-	
misure disciplinari	provvedimento disciplinare/procedimenti disciplinari/azione disciplinare	3	<jrA'At t>dybyp	tdAbyr t>dybyp	1
inchiesta preliminare	indagine preliminare	1	AltHqyqAt Al>wlyp	AlfHS Al>wly	1
opinione pubblica	dibattito pubblico/attenzione pubblica/risonanza pubblica	3	Alr>y AlEAm	AljmAhyr AlEAmp	1
incitamento alla lotta settaria	istigazione alla lotta settaria	1	AltHryD ElY Alftnp AlTA}fyp	<vArp AlnErAt AlTA}fyp	1
ordine pubblico	--		AlnZAm AlEAm	AlmSlHp AlEAmp/AlSAlH AlEAm	2
morale pubblica	--	1	Al>xlAq AlEAmp	Al|dAb AlEAmp	1
ambiente di lavoro	luogo di lavoro	1	by}p AlEml	mwAqE AlEml/>mAkn AlEml	2
direzione criminale investigativa	dipartimento investigativo criminale/polizia investigativa criminale	2	<dArp AltHqyqAt AljnA}yp	$rTp AltHqyqAt AljnA}yp/dArp AlmbAHv AljnA}yp	2
obbligo scolastico	istruzione scolastica	1	AltElym Al<lzAmY	-	
sfruttamento sessuale	abuso sessuale	1	AlAstglAl Aljnsy	AlAEtdA' Aljnsy/Al<sA'p Aljnsyp	2
assicurazione sociale	previdenza sociale	1	AlDmAn AlAjtmAEy	AlHmAyp AlAjtmAEyp	1
stupro coniugale	abuso coniugale/violenza coniugale	2	AlAgtSAb Alzwjy	AlEnf Alzwjy	1
inchiesta indipendente	indagine indipendente	1	tHqyq mstql	-	
perizia legale	medico legale/referto medico-legale	2	AlTb Al$rEy	Altqryr Al$rEy	1
sentenza definitiva	giudizio definitivo	1	Hkm nhA}y	Al<dAnp AlnhA}y/qrAr nhA}Y	2

disfunzione nel sistema giudiziario	carenze nel sistema giudiziario/inefficacia del sistema giudiziario/debolezza del sistema giudiziario	3	qSwr fy AlnZAm AlqDA}y	Edm fEAlyp AlnZAm AlqDA}y/tqAEs AlnZAm AlqDA}y	2
revisione giudiziaria	riesame giudiziario	1	AlmrAjEp AlqDA}yp	<EAdp nZr qDA}yp	1
camera d'appello	collegio d'appello	1	dA}rp AlAst}nAf	grfp AlAst}nAf	1
sicurezza personale	incolumità personale	1	Al>mn Al$xSy	AlslAmp Al$xSyp	1
protezione dei testimoni	incolumità dei testimoni	1	HmAyp Al$hwd	msAndp Al$hwd	1
conflitto d’interesse	--		tDArb AlmSAlH	AlSrAEAt ElY AlmSAlH	1
annullamento di un contratto	cessazione di un contratto/rescissione di un contratto	2	fsx Eqd	<nhA' Eqd/AlgA' Eqd	2
ordine di confisca	decisione di confisca/ordinanza di confisca/provvedimento di confisca	3	>mr AlmSAdrp	qrAr AlmSAdrp	1
parità delle opportunità	eguaglianza delle opportunità	1	tkAf& AlfrS	tsAwy AlfrS	1
scrutinio segreto	voto segreto	1	AlAqtrAE Alsry	AltSwyt Alsry	1
diritto sindacale	--		AlHryp AlnqAbyp	Hq AltnZym AlnqAby	1
controllo giudiziario	--		rqAbp qDA}yp	tHryAt qDA}yp/mrAjEp qDA}yp	2
inserimento sociale	inclusione sociale	1	Al<dmAj AlAjtmAEy	Al<$rAk AlAjtmAEy	1
riduzione di pena	alleggerimento della pena/commutazione della pena	2	txfyf llEqwbp	<bdAl AlEqwbp	1
soluzione delle controversie	composizione delle controversie	1	tswyp AlmnAzEAt	Hl AlmnAzEAt	1
rappresentante legale	avvocato rappresentante	1	Almmvl AlqAnwny	AlwSy AlqAnwny	1
protezione della maternità	tutela della maternità	1	HmAyp Al>mwmp	rEAyp Al>mwmp	1
punizioni corporali	pene corporali	1	AlEqwbAt Albdnyp	-	
mortalità materna	decessi materni	1	AlwfyAt byn Al>mhAt	-	
reintroduzione della legge	reiterazione della legge	1	<EAdp AlEml bqAnwn	-	
causa civile	azione civile/procedimento civile	2	AldEwY Almdnyp	$kwY mdnyp/AlqDyp Almdnyp	2
divulgazione di segreti di stato	rivelazione di segreti di stato	1	<f$A' >srAr Aldwlp	tsryb >srAr Aldwlp	1
protocollo d’intesa	memorandum d’intesa	1	m*krp AltfAhm	-	
lacune legislative	scappatoie legislative	1	AlvgrAt Alt$ryEyp	Alfjwp fy Alt$ryEAt	1
rilascio anticipato	proscioglimento anticipato/liberazione anticipata	2	Al<frAj Almbkr	-	
disposizioni della sharia	legge della sharia/codice della sharia/legislazione della sharia	3	>HkAm Al$ryEp	qwAEd Al$ryEp/tEAlym Al$ryEp/qwAnyn Al$ryEp	3
composizione amichevole	accordo amichevole/soluzione amichevole	2	tswyp wdyp	Hl wdy	1

pregiudizio politico	parzialità politica	1	AlAnHyAz AlsyAsy	-	
reclutamento dei giudici	designazione dei giudici/nomina dei giudici	2	twZyf AlqDAp	tEyyn AlqDAp	1
notifica di sgombero	ordinanza di sgombero	1	<$EAr bAl<xlA'	<n*Ar bAl<xlA'/<$EArAF bAl<xlA'/>mr Al<xlA'	3
popolazioni native	comunità native	1	AlskAn Al>Slyyn	Al$Ewb Al>Slyp	1
conflitto armato	--		AlnzAE AlmslH	AlSrAE AlmslH	1
associazione a delinquere	--		Alt|mr AljnA}y	AlmxTT AljnA}y	1
manifestazioni antigovernative	marce antigovernative/proteste antigovernative	2	AlmZAhrAt AlmEArDp llHkwmp	AlAHtjAjAt AlmnAhDp llHkwmp/AlmsyrAt AlmnAhDp llHkwmp	2
certificato di nascita	--		$hAdAt AlmylAd	w}A}q AlmylAd	1
racconti di testimoni	affermazioni di testimoni	1	>qwAl Al$hwd	$hAdAt Al$hwd/rwAyAt Al$hwd	2
parità di trattamento	eguaglianza di trattamento	1	AlmsAwAp fy AlmEAmlp	-	
infortuni sul lavoro	incidenti sul lavoro	1	AlHwAdv Almhnyp	AlmxATr Almhnyp/ Al<ySAbAt Almhnyp	2
seggio vacante	posto vacante	1	AlmqEd Al$Agr	AlmnSb Al$Agr	1
stato di diritto	primato del diritto	1	syAdp AlqAnwn	Hkm AlqAnwn	1
istruzione primaria	educazione primaria/insegnamento primario	2	AltElym Al>sAsy	-	
riabilitazione professionale	reinserimento professionale/riadattamento professionale	2	AEAdp Alt>hyl Almhny	-	
sviluppo della carriera	avanzamento di carriera	1	AltTwr AlwZyfy	Altqdm AlwZyfy	1
notifica di licenziamento	preavviso di licenziamento	1	<xTAr bAlfSl	An*ArA bAlfSl	1
condizioni di impiego	--		$rwT >stxdAm	>wDAE Al>stxdAm	1
processo decisionale	potere decisionale	1	SnE AlqrAr	AtxA* AlqrAr	1
relazioni familiari	--		AlElAqAt Al>sryp	AlrwAbT Al>sryp	1
persone giuridiche	--		Al$xSyAt AlAEtbAryp	Alhy}At AlAEtbAryp	1
tecniche di interrogatorio	procedure di interrogatorio	1	wsA}l AstjwAb	>sAlyb AstjwAb	1
diniego della cittadinanza	revoca della cittadinanza	1	Altjryd mn Aljnsyp	AlHrmAn mn Aljnsyp	1
misure organizzative	--		AltdAbyr AltnZymyp	trtybAt tnZymyp/qwAEd tnZymyp/<jrA'At tnZymyp	3
mandato giudiziario	autorizzazione giudiziaria	1	<*n qDA}y	-	
reati amministrativi	crimini amministrativi/irregolarità amministrative/illeciti amministrativi	3	mxAlfAt <dAryp	jrA}m <dAryp/thm <dAryp	2
accesso all'istruzione	diritto all'istruzione	1	AlHSwl ElY AltElym	AlAltHAq bm&ssAt AltElym	1
scadenza della condanna	-	1	AnqDA' mdp AlHkm	<tmAm mdp AlHkm	1

omicidi politici	uccisioni politiche	1	qtl syAsy	AgtyAl syAsy	1
mancanza di sicurezza	precarie condizioni di sicurezza	1	AnEdAm Al>mn	AftqAr Al>mn/fqdAn Al>mn/nqS fy Al>mn	3
attentati suicidi	attacchi suicidi	1	hjmAt AntHAryp	EmlyAt AntHAryp/tfjyrAt AntHAryp	2
brogli elettorali	frodi elettorali/ irregolarità elettorali/ manipolazioni elettorali	3	tzwyr AlAntxAbAt	AltlAEb bntA}j AlAntxAbAt/AlAnthAkAt AlAntxAbyp/mxAlfAt AntxAbyp	3
persone eminenti	eminenti figure	1	$xSyAt bArzp	-	
bozza di costituzione	--		mswdp Aldstwr	m$rwE Aldstwr	1
sistema di giustizia militare	codice di giustizia militare	1	qAnwn Al>HkAm AlEskryp	nZAm AlqDA' AlEskry/qAnwn AlqDA' AlEskry/'qAnwn AljnAyAt AlEskryp	3
soffocare l'opposizione	--		kbt AlmEArDp	tkmym AlmEArDp	1
assemblea nazionale	associazione nazionale	1	AljmEyp AlwTnyp	Almjls AlwTny	1
organizzazioni cooperative	strutture cooperative	1	AlmnZmAt AltEAwnyp	AlhyAkl AltEAwnyp	1
termine ultimo	--		Al>Tr Alzmnyp	Almhlp Alzmnyp/Almdd Alzmnyp	2
misure di ispezione	sistemi di ispezione	1	nZAm tfty$	EmlyAt Altfty$	1
certificato medico di attitudine	esame medico di attitudine	1	tqryr AllyAqp	AlfHS AlTbY llyAqp	1
consulenza giuridica	--		Alm$wrp AlqAnwnyp	AlmsAEdp AlqAnwnyp/AlxdmAt AlqAnwnyp	2
revoca di immunità	sospensione dell'immunità	1	rfE AlHSAnp	<nhA' AlHSAnp	1
misure provvisorie	--		tdAbyr m&qtp	<jrA'At m&qtp	1
rientro di profughi	rimpatrio dei profughi	1	<EAdp AllAj}yn	Ewdp AllAj}yn	1
distruzione di proprietà pubblica	danneggiamento di proprietà pubblica	1	<tlAf AlmmtlkAt AlEAmp	<lHAq AlDrr bAlmmtlkAt AlEAmp	1
tasso di disoccupazione	livello di disoccupazione	1	mEdl AlbTAlp	nsbp AlbTAlp	1
ricongiungimento familiare	riunificazione familiare	1	lm $ml >srp	jmE $ml Al>srp	1
quotidiano a capitale privato	giornale a capitale privato	1	SHyfp xASp	-	
piano di azione	programma di azione	1	xTp Eml	brnAmj AlEml	1
periodo di preavviso	tempo di preavviso	1	mhlp Al<n*Ar	-	
scambio di informazioni	diffusione di informazioni /divulgazione di informazioni	2	tbAdl AlmElwmAt	m$Arkp AlmElwmAt	1
misure correttive	provvedimenti correttivi	1	tdAbyr tSHyHyp	AjrA'At tSHyHyp	1
metodi di pagamento	sistemi di pagamento/modi di pagamento/modalità di pagamento	3	>sAlyb AldfE	TrA}q dfE	1

organi sussidiari	organismi sussidiari	1	Alhy}At AlfrEyp	-	
periodo di prescrizione	termine di prescrizione	1	mdp tqAdm	ftrp tqAdm	1
contabilità nazionale	bilancio nazionale	1	AlHsAbAt Alqwmyp	-	
pericolo pubblico	emergenza pubblica	1	AlTwAr} AlEAmp	AlxTr AlEAm	1
comunità internazionale	---		AlmjtmE Aldwly	jhAt dwlyp	1
salario base	paga base	1	Al>jr Al>sAsy	-	
imprese multinazionali	società multinazionali	1	Almn$|t mtEddp Aljnsyp	Alm&ssAt mtEddp Aljnsyp/$rkAt mtEddp AljnsyAt	2
livello di vita	standard di vita/tenore di vita/condizioni di vita	3	mstwY mEy$p	Zrwf AlmEy$p	1
misure d’austerità	provvedimenti d’austerità	1	<jrA'At Altq$f	tdAbyr Altq$f	1
contaminazione ambientale	inquinamento ambientale	1	tlwv Alby}p	tdmyr Alby}p	1
atti terroristici	azioni terroristiche/attività terroristiche	2	>EmAl <rhAbyp	>n$Tp <rhAbyp	1
codice elettorale	legge elettorale/legislazione elettorale/normativa elettorale	3	qAnwn AlAntxAbAt	qwAEd AlAntxAbAt	1
organizzazione studentesca	associazione studentesca/unione studentesca	2	mnZmp TlAbyp	AtHAd AlTlbp	1
personale carcerario	guardie carcerarie/funzionari carcerari	2	ms&wly Alsjwn	mwZfy Alsjwn	1
alta autorità morale	alta levatura morale	1	>ElY AlSfAt Alxlqyp	-	
stato membro	paese membro	1	Aldwl Al>TrAf'	-	
stereotipi sessisti	pregiudizi sessisti	1	AlqwAlb AlnmTyp ؛dwAr Aljnsyn	-	
azione coercitiva	metodi coercitivi/pratiche coercitive/mezzi coercitivi/misure coercitive	4	tdAbyr AlqmE	>EmAl AlqmE	1
minoranza religiosa	gruppi religiosi	1	>qlyp dynyp	AlTwA}f Aldynyp/jmAEAt dynyp	2
atti di violenza	azioni di violenza/episodi di violenza	2	>EmAl AlEnf	>HdAv Enf/HwAdv AlEnf/>$kAl AlEnf/>fEAl AlEnf	4
incitamento alla violenza	istigazione alla violenza	1	AltHryD Ely AlEnf	t$jyE AlEnf	1
tutela dei diritti umani	protezione dei diritti umani/difesa dei diritti umani	2	HmAyp Hqwq AlAnsAn	tEzyz Hqwq Al<nsAn	1
organizzazioni dei diritti umani	associazione dei diritti umani	1	mnZmp Hqwq Al<nsAn	jmEyp Hqwq Al<nsAn	1
centro di detenzione	struttura di detenzione/campo di detenzione/luogo di detenzione	3	mrAkz AlAHtjAz	>mAkn AlAHtjAz	1
coordinazione internazionale	coordinamento internazionale	1	tnsyq dwly	tEAwn dwly	1
direttore generale	amministratore generale	1	mdyr EAm	-	
autorità locale	enti locali/amministrazione locale	2	slTp mHlyp	<dArp mHlyp	1

standard internazionali	regolamento internazionale/norme internazionali	2	mEAyyr dwlyp	mstwyAt dwlyp/mqAyys dwlyp	2
personale di sicurezza	funzionari di sicurezza/agenti della sicurezza	2	ms}wl Al>mn	>frAd Al>mn/rjAl Al>mn	2
formazione professionale	--		Alt>hyl Almhny	Altdryb Almhny/Altwjyh AlmhnY	2
manifestazione pacifica	protesta pacifica	1	msyrp slmyp	mZAhrp slmyp/AHtjAj slmy	2
lotta alla violenza	contrasto alla violenza		mHArbp AlEnf	mEAljp AlEnf/mkAfHp AlEnf	2
dispositivo di protezione individuale	attrezzature di protezione individuale/ equipaggiamento di protezione individuale	2	mEdAt AlwqAyp Al$xSyp	-	
applicazione delle raccomandazioni	esecuzione delle raccomandazioni/attuazione delle raccomandazioni	2	tnfy* AltwSyAt	tTbyq twSyAt	1
segnalazioni di tortura	episodi di tortura/denunce di tortura/accuse di tortura/casi di tortura	4	HAlp tE*yb	HwAdv AltE*yb/AdEA'At AltE*yb/>fEAl tE*yb/$kAwY AltE*yb/mmArsp AltE*yb	5
molestia sessuale	aggressione sessuale	1	tHr$ jnsy	AEtdA' jnsy	1
regime militare	governo militare/giunta militare	2	nZAm Eskry	AlHkm AlEskry/AlqyAdp AlEskryp	2
protezione dei civili	tutela dei civili	1	HmAyp Almdnyyn	-	
sistema penitenziario	regime penitenziario	1	nZAm Alsjwn	-	
posto di lavoro	opportunità di lavoro	1	frSp Eml	mkAn AlEml	1
misure legislative	atti legislativi	1	AltdAbyr Alt$ryEyp	Al<jrA'At Alt$ryEyp	1
carcere ai lavori forzati	ergastolo ai lavori forzati/reclusione ai lavori forzati	2	Alsjn mE Al>$gAl Al$Aqp	>HkAm bAl>$gAl Al$Aqp	1
crimine di omicidio	reato di omicidio	1	jrymp qtl	Emlyp Alqtl	1
proclamare lo stato di emergenza	imporre lo stato di emergenza	1	<ElAn HAlp AlTwrA}	frD HAlp AlTwAr}	1
permesso di lavoro	--		tSryH Eml	A*n AlEml	1
sentenze con sospensione della pena	abolizione della pena/moratoria sulla pena	2	>HkAm mE wqf Altnfy*	-	
pagamento di un'ammenda	--		dfE grAmp	frD grAmp	1
pubblicazione di notizie false	diffusione di notizie false/trasmissione di notizie false	2	n$r >xbAr kA*bp	<*AEp >nbA' kA*bp	1
casellario penale	fedina penale/precedenti penali	2	sjl jnA}y	swAbq jnA}yp	1
attivisti dei diritti umani	difensori dei diritti umani	1	nA$Tw Hqwq Al<nsAn	-	
divieto di discriminazione	eliminazione di discriminazione	1	HZr Altmyyz	AlqDA' ElY Altmyyz/mnE Altmyyz /<zAl_p Altmyyz	3
paralisi politica	impasse politica	1	jmwd syAsy	Al$ll AlsyAsy/AnsdAd AlTryq AlsyAsy/m>zq syAsy/AnsdAd Al>fq AlsyAsy	4
prassi discriminatorie	pratiche discriminatorie	1	AlmmArsAt Altmyyzyp	-	

tratta di esseri umani	traffico di esseri umani	1	AlAtjAr bAlb$r	-	
determinazione della pena	-		tqryr AlEqwbp	twqE Eqwbp	1
scontare una pena	-		<tmAm mdp AlHkm	qDA' mdp AlHkm/tnfy* mdp AlHkm/t>dyp mdp AlHkm/AstyfA' mdp AlHkm	4
sistema di licenza	regolamento di licenza/regime di licenza	2	nZAm lltrxyS	--	
revoca di licenza	ritiro di licenza	1	sHb AltrxyS	tjmyd trxyS/wqf trxyS	2
agevolazioni fiscali	incentivi fiscali/sgravi fiscali	2	HwAfz Drybyp	mzAyA Drybyp	1
dati base	banche dati	1	qwAEd AlbyAnAt	--	
rapina aggravata	furto aggravato	1	AlsTw AlmslH	srqp mslHp	1
rappresentante diplomatico	agente diplomatico/ funzionario diplomatico	2	mmvl dblwmAsy	mwZf dblwmAsy/EDw Alslk AldblwmAsy	2
mezzi di comunicazione	-		wsA}l Al<ElAm	wsA}T Al<ElAm	1
effetto retroattivo	applicazione retroattiva	1	>vr rjEy	--	
consiglio elettorale	commissione elettorale/comitato elettorale/collegio elettorale	3	mjls AlAntxAbAt	hy}p AlAntxAbAt/ljnp AlAntxAbAt/mfwDyp AlAntxAbAt	3
tecniche di interrogatorio rinforzate	procedure di interrogatorio rinforzate/ metodi di interrogatorio rinforzate	2	>sAlyb AstjwAb m$ddp	AdwAt AstjwAb m$ddp/ Trq AstjwAb m$ddp	2
violazioni dei diritti umani	abusi dei diritti umani	1	AnthAk Hqwq AlAnsAn	AEtdA'At ElY Hqwq Al<nsAn	1
documento d'entità	carta d'entità	1	wvA}q Alhwyp	bTAqAt Alhwyp	1
disturbo dell'ordine pubblico	-	1	tEkyr Sfw AlnZAm AlEAm	AxlAl bAlnZAm AlEAm/zEzEp AlnZAm AlEAm	2
totale	298		270		

6.3.1.2. sinonimia parziale con sostituzione di un modificatore

Termine di base italiano	Varianti del Termine di base italiano	n	Termine di base arabo	Varianti del Termine di base arabo	n
pena di morte	pena capitale	1	Hkm Al<EdAm		
violenza domestica	violenza all'interno della famiglia/violenza familiare	2	AlEnf Al>sry	AlEnf fy <TAr Al>srp/'AlEnf dAxl Al>srp /AlEnf fy mHyT Al>srp	3
funzionario pubblico	funzionario statale/funzionario governativo	2	mwZf Emwmy	mwZfy AlHkwmp/mwZfy Aldwlp/mwZfyn fy Aldwlp	3
minoranza etnica	minoranza razziale	1	>qlyAt Erqyp	>qlyAt <vnyp	1
detenzione arbitraria	detenzione senza accusa/detenzione illegale/detenzione sommaria	3	AEtqAl tEsfy	AEtqAl gyr qAnwny	1
rimpatrio forzato	rimpatrio non spontaneo	1	AbEAd qsry	<bEAd gyr $rEy	1
autorità competente	autorità incaricata/autorità appropriata/autorità qualificata/autorità	6	AlslTp AlmxtSp	AlslTp Alms}wlp/AlslTp AlHkwmyp/AlslTp AlEAmp	3

	investita/autorità decisionale/autorità pubblica				
prigioniero di guerra	soldato prigioniero	1	>srY AlHrb		
legge sui crimini internazionali	legge sui reati internazionali	1	qAnwn AljrA}m Aldwlyp	AlqAnwn AljnA}y Aldwly/qAnwn mHkmp AljnAyAt Aldwlyp/qAnwn AlmHAkm Aldwlyp	3
difensore civico	difensore pubblico	1	AlmdAfE AlEAm		
diritti umani	diritti delle persone/diritto umanitario	2	Hqwq Al<nsAn	Hqwq Al>$xAS	1
agente di polizia	agente della sicurezza	1	DAbT $rTp	DAbT >mn	1
libertà di espressione	libertà di parola/libertà di pensiero	2	Hryp AltEbyr	Hryp AlklAm/Hryp Alr>y	2
legislazione anti-terrorismo	-		qAnwn mkAfHp Al<rhAb	qAnwn mnE Al<rhAb/qAnwn qmE Al<rhAb	2
procedimento giudiziario	procedimento penale	1	Al<jrA'At AlqDA}yp	<jrA'At AlmHAkmp/<jrA'At jnA}yp	2
ministro della giustizia	-		wzyr AlEdl	wzyr AlqAnwn/wzyr Al$&wn AlqAnwnyp	2
codice penale	-		AlqAnwn AljnA}y	qAnwn AlEqwbAt	1
servizio militare	-		Alxdmp AlEskryp	Alxdmp fy Aljy$	1
discriminazione razziale	discriminazione etnica	1	Altmyyz AlEnSry	Altmyyz ElY >sAs AlErq	1
mandato di arresto	mandato di cattura	1	>mr bAlqbD	>mr AlsyTrp/>wAmr twqyf	2
richiesta di estradizione	richiesta di consegna		Tlb Altslym	Tlb tqdym	1
diritti dei minori	diritti dell'infanzia/diritti dei bambini/diritti dei minorenni/diritti del fanciullo	4	Hqwq AlTfl	Hqwq AlqSr	1
libertà di stampa	libertà dei mezzi d'informazione /libertà degli organi d'informazione	2	Hryp AlSHAfp	Hryp wsA}l Al<ElAm	1
disposizioni di legge	disposizioni legislative	1	Al>HkAm AlqAnwnyp	AlAHkAm Al$ryEyp	1
congedo pagato	congedo retribuito	1	<jAzp mdfwEp Al>jr		
lavoratori migranti	lavoratore emigrante/lavoratori immigrati	2	AlEmAl AlmhAjryn	AlEmAl AlwAfdyn/AlEmAl Al>jAnb/AlEmAl Al>jAnb AlwAfdyn	3
lavoro minorile	lavoro dei bambini/lavoro infantile	2	Eml Al>TfAl		1
giudice inquirente	giudice istruttore/giudice titolare	2	qADy AltHqyq		
legislazione nazionale	legislazione interna	1	qAnwn wTny	AlqAnwn AlmHly / qnwn dAxly	2
matrimonio omosessuale	matrimonio tra persone dello stesso sesso	1	zwAj mvly	>zwAj mn Aljns AlwAHd/AlzwAj byn >frAd mn nfs Aljns/	2
sistema penitenziario	sistema carcerario/sistema di prigioni	2	nZAm Alsjwn	>nZmp AHtjAz	1
emendamento alla legge	emendamenti al codice	1	tEdyl qAnwn	tEdylAt t$ryEyp	1
risarcimento economico	risarcimento finanziario/risarcimento monetario/risarcimento in denaro	2	tEwyD mAlY		
traffico di droga	traffico di stupefacenti		tjArp AlmxdrAt		
codice di condotta	codice di comportamento/codici comportamentali	2	qwAEd Alslwk		
polizia di rapido intervento			$rTp Altdxl AlsryE	$rTp Alrd AlsryE/$rTp AlTrq AlsryEp'	2
oltraggio a pubblica autorità	oltraggio a pubblico ufficiale	1	AltEdy ElY ms&wl EAm	AltEdy ElY mwZf EAm	1
omicidio volontario	omicidio premeditato/omicidio aggravato/ omicidio intenzionale	3	Alqtl AlEmd	Alqtl mE sbq Al<SrAr wAltrSd	1

omicidio colposo	omicidio involontario	1	Alqtl AlxT>	Alqtl gyr AlmtEmd/ Alqtl AlADTrAry	2
tentato omicidio			Al$rwE fy Alqtl	Al$rwE fy Emlyp AgtyAl	1
tratta di bambini	tratta di minori	1	AlAtjAr fy Al>TfAl		
decesso in custodia	decesso in detenzione	1	wfAp fy AlHjz	wfAp fy Alsjn/ wfAp fy AlznzAnp	2
consiglio supremo della magistratura	consiglio supremo giudiziario/consiglio superiore della magistratura	2	Almjls Al>ElY llqDA'		
sentenza con sospensione della pena			>HkAm mE wqf Altnfy*	>HkAm gyr nAf*p	1
appartenenza a gruppo criminale	appartenenza a un'organizzazione criminale/appartenenza a un gruppo illegale/appartenenza a banda criminale/appartenenza ad associazione criminale	4	AlAntmA' <lY jmAEp <jrAmyp	AlAntmA' <lY jmAEp gyr qAnwnyp /AlAntmA' <lY mnZmp AjrAmyp/AlAntmA' <lY tnZym <jrAmy	3
detenzione preprocessuale	detenzione preliminare/detenzione preventiva	2	AlAHtjAz AlsAbq llmHAkmp	AlAHtjAz AlAHtyATy	1
abuso d'ufficio	abuso di autorità/abuso di potere	2	<sA'p AstxdAm AlslTp	<sA'p AstglAl AlslTp/<sA'p AstEmAl AlslTp	2
rapporti sessuali illeciti	rapporti sessuali extraconiugali	1	ElAqAt jnsyp gyr $rEyp	ElAqAt jnsyp mE gyr >zwAjhn	1
immunità giudiziaria	immunità processuale/immunità penale	2	HSAnpF mn AlmqADAp AljnA}yp	HSAnp mn AlmlAHqp AlqDA}yp	1
indipendenza della magistratura	indipendenza dell’autorità giudiziaria/l'indipendenza degli organi giudiziari	2	AstqlAl AlqDA'	AstqlAl AlmHkmp	1
personalità giuridica	--		Al$xSyp AlqAnwnyp	Al$xSyp Al>EtbAryp	1
salute professionale	salute sul lavoro/ salute dei lavoratori	2	AlSHp Almhnyp	SHp AlEmAl/AlSHp >vnA' AlEml	2
malattia professionale	malattie sul lavoro	1	Al>mrAD Almhnyp		
sviluppi legislativi	sviluppi giuridici/sviluppi legali	2	AltTwrAt AlqAnwnyp	AltTwrAt Alt$ryEyp	1
direzione criminale investigativa			<dArp AltHqyqAt AljnA}yp	<dArp AlbHv AljnA}y	1
congedo di maternità	congedo post-natale	1	<jAzp AlwDE	<jAzp Al>mwmp	1
organizzazioni umanitari	organizzazioni volontarie/organizzazioni di volontariato/organizzazioni assistenziali/organizzazioni di aiuti/organizzazioni volontarie	4	AlmnZmAt AltTwEyp	mnZmAt Al<gAvp Al<nsAnyp /AlmnZmAt AlEAmlp fy mjAl AlmsAEdAt Al<nsAnyp	2
cittadini stranieri	-		AlmwATnyn Al>jAnb	gyr AlmwATnyn	1
sentenza definitiva	sentenza finale	1	Hkm nhA}y	Hkm bAt	1
testimone dell'accusa	-		$hwd AlAdEA'	$hwd AlAthAm	1
testimoni della difesa	testimoni a discarico		$hwd AldfAE	$hwd Alnfy/$hwd Altbr}p	2
revisione giudiziario	-		mrAjEp qDA}yp	mrAjEp qAnwnyp	1
sicurezza nazionale	sicurezza dello Stato, sicurezza interna	2	Al>mn AlwTny	Al>mn Alqwmy />mn Aldwlp	2
diritto alla difesa			AlHq fy AldfAE	Hqhm fy AlAstEAnp bmHAmyn	1
sviluppo sostenibile	sviluppo durevole	1	Altnmyp AlmstdAmp		
proventi del reato	proventi del crimine	1	EA}dAt Aljrymp	AlEA}dAt Almt>typ mn Al>fEAl Almjr~mp	1
infanticidio femminile	Infanticidio delle figlie	1	>d Al<nAv	w>d AlbnAt	1

diritto sindacale	diritto di organizzazione /diritto di contrattazione collettiva	2	AlHq fy AltnZym	Hq AltnZym AlnqAby/AlHq AlnqAby	2
controllo giudiziario	controllo giurisdizionale	1	rqAbp qDA}yp.		
pregiudizi etnici	pregiudizi sessisti/pregiudizi razziali	2	AlAnHyAz AlErqy	AnHyAzhA AlEnSry	1
autore di reato	autore di crimine	1	mrtkb Aljrymp		
immigrazione clandestina	immigrazione illegale/immigrazione irregolare	2	Alhjrp gyr Al$rEyp	Alhjrp Almsttrp /Alhjrp gyr AlqAnwnyp	2
punizioni corporali	punizioni fisiche	1	AlEqwbAt Albdnyp	AlEqwbAt Aljsdyp	1
legge sulla sedizione	-		qAnwn AlESyAn	qAnwn Alftnp	1
sanzioni penali	-		EqwbAt jnA}yp	EqwbAt jzA}yp	1
cospirazione criminale	-		Alt|mr AljnA}y	Alt|mr lArtkAb jrymp	1
diritti dell'imputato	diritti dell’accusato /diritti della difesa	2	Hqwq Almthm	Hqwq AlmdEY Elyh /Hqwq AldfAE	2
accusa inventata	accusa costruita/ accusa falsa	2	thmp mlfqp	thm mzEwmp	1
lacune legislative	lacune normative/lacune del sistema giudiziario	2	AlvgrAt Alt$ryEyp	AlvgrAt AlqAnwnyp	1
perquisizioni corporali	perquisizioni a nudo	1	Altfty$ Aljsdy	Altfty$ Al*Aty	1
rilascio anticipato	rilascio con la condizionale	1	Al<frAj Almbkr	<frAj m$rwT	1
organizzazione messa al bando	organizzazione posta al bando /organizzazione vietata	2	mnZmp mHZwrp	mnZmp gyr mrxSp /mnZmp gyr qAnwnyp	2
forze filogovernative			AlqwAt AlmwAlyp llHkwmp	AlqwAt Alm&ydp llHkwmp /AlqwAt AltAbEp llnZAm	2
diritto agli indennizzi	diritto alla riparazione giudiziaria/diritto a un risarcimento	2	AlHq fy AltEwyDAt	AlHq fy Al<nSAf	1
diritto all'alloggio	diritto a un'abitazione/diritti abitativi/diritti di locazione	3	Hqwq Alskn		
popolazioni native	popolazioni indigene	1	AlskAn Al>Slyyn		
uguaglianza di genere	uguaglianza tra i sessi/uguaglianze tra uomini e donne	2	AlmsAwAp byn Aljnsyn	AlmsAwAp byn Almr>p wAlrjl/AlmsAwAp AlqA}m ElY nwE Aljns/AlmsAwAp fy AlnwE AlAjtmAEy	3
salute riproduttiva			AlSHp Al<njAbyp	AlSHp AltnAslyp	1
censura preventiva alla stampa	censura prima della pubblicazione	1	AlrqAbp Almsbqp ElY AlSHf		
possesso illegale di armi	possesso illecito di armi	1	HyAzp >slHp b$kl gyr qAnwny	HyAzp >slHp gyr m$rwEp	1
associazione a delinquere	associazione criminale	1	Alt|mr AljnA}y	Alt|mr lArtkAb jnAyp	1
manifestazioni antigovernative	manifestazioni contro il governo	1	AlmZAhrAt AlmEArDp llHkwmp	AlmZAhrAt AlmnAhDp llHkwmp	1
rapina a mano armata	rapina aggravata	1	AlsTw AlmslH	AlsTw Alm$dd	1
diritto di appello	diritto di ricorso	1	AlHq fy AlAst}nAf	AlHq fy AlTEn/AlHq fy rfE dEAwY Ast$kAl	2
camera dei deputati	camera dei rappresentanti	1	mjls AlnwAb	mjls AlEmwm	1
libertà di associazione	libertà di riunione	1	Hryp tkwyn AljmEyAt	Hryp AltjmE	1
diritto alla riservatezza	diritto alla vita privata	1	Hq AlxSwSyp		
infortuni sul lavoro	-		<SAbAt mhnyp	<SAbAt AlEml	1
persone disabili	persone portatrici di handicap	1	Al>$xAS *wY Al<EAqp	Al>$xAS AlmEAqyn	1

divieto totale	divieto assoluto	1	AlHZr Al$Aml	AlHZr Alkly/AlHZr AlmTlq	2
istruzione primaria	istruzione di base	1	AltElym Al>sAsy	AltElym AlAbtdA}y	1
confisca di beni	confisca di proprietà	1	mSAdrp >Swl	mSAdrp mmtlkAt	1
effettivo esercizio	-		AlmmArsp AlfEAlp	AlmmArsp AlEmlyp	1
condizioni di impiego	condizioni di arruolamento	1	$rwT >stxdAm	$rwT AltwZyf /$rwT AlEml	2
contratto di lavoro	contratto di impiego	1	Eqd Eml	Eqwd AltwZyf	1
certificato di competenza	certificato di capacità	1	$hAdp kfA'p		
competenza della corte	competenza del tribunale	1	AxtSAS AlmHkmp		
applicazione della legge	applicazione della normativa/applicazione della legislazione	2	tTbyq qwAnyn	tTbyq t$ryE	1
parti del conflitto	parti belligeranti/parti in lotta	1	>TrAf AlnzAE	>TrAf AlqtAl/>TrAf AlSrAE	2
protezione del posto di lavoro	protezione del reddito dei lavoratori	1	AlHmAyp AlwZyfyp	HmAyp AlEml	1
relazioni familiari	relazioni domestiche	1	AlElAqAt Al>sryp	AlElAqAt Alzwjyp	1
profilazione etnica	profilazione razziale	1	AsthdAf Erqy	AsthdAf EnSry	1
indagini preliminari	indagini basilari/indagini di base	2	AltHqyqAt Al>wlyp	AltHqyqAt Al>sAsyp	1
stupro di gruppo	stupro di massa	1	AlAgtSAb AljmAEy		
sterilizzazione forzata	Sterilizzazione illegale/sterilizzazione senza il consenso pieno/sterilizzazione coercitiva	3	AltEqym Alqsry	AltEqym gyr AlqAnwny	1
accesso all'istruzione	accesso all'insegnamento/accesso all'educazione	2	AlHSwl ElY AltElym		
destabilizzazione del regime	destabilizzazione del potere dello stato/destabilizzazione del governo	2	zEzEp AlHkm	zEzEp AstqrAr AlnZAm/zEzEp slTp Aldwlp	2
presidente uscente	ex presidente	1	Alr}ys AlmnSrf	Alr}ys Almnthyp wlAyth/Alr}ys AlsAbq	2
prestazioni in denari	prestazioni finanziarie	1	Al<EAnAt Alnqdyp		
agenzie di collocamento	agenzie di lavoro	1	wkAlAt Alt$gyl	wkAlAt AltwZyf /wkAlAt AlEml	2
termine ultimo	termini temporali/termini di tempo stabiliti/termini previsti	3	Al>Tr Alzmnyp		
misure provvisorie	misure ad interim/misure cautelari/misure temporanee	3	tdAbyr m&qtp		
principio del non-refoulement	principio del non rimpatrio forzato	1	mbd> HZr Al<EAdp Alqsryp		
contratto di lavoro contemporaneo	contratto a tempo determinato	1	Eqwd Eml lmdp mHddp	Eqwd Eml lmhmp mHddp/Eqwd Eml lmdp vlAv snwAt gyr qAblp lltjdyd	2
età lavorativa	età attiva	1	sn AlEml		
assicurazione obbligatoria			Alt>myn Al<lzAmy	Alt>myn Al<jbAry	1
servizi pubblici			AlxdmAt AlEAmp	AlxdmAt AlHkwmyp	1
misure correttiva			tdAbyr tSHyHyp	tdAbyr AntSAf/tdAbyr ElAjyp	2
fabbricazione illecita stupefacenti	fabbricazione clandestina di tali stupefacenti	1	SnE AlmxdrAt bSwrp gyr m$rwEp		
comunità internazionale			AlmjtmE Aldwly	AlmjtmE AlEAlmy	1
salario base	salario forfettario	1	Al>jr Al>sAsy	Al>jr Al<jmAly	1
livello di vita			mstwY mEy$p	mstwY AlHyAp	1
crisi economica	crisi finanziaria	1	>zmp AqtSAdyp	>zmp mAlyp	1

sciopero generale	sciopero nazionale	1	<DrAb EAm	<DrAb wTny	1
coinvolgimento attivo			m$Arkp n$Tp	m$Arkp fEAlp	1
permesso di soggiorno	permessi di residenza	1	tSAryH <qAmp		1
bilancio nazionale	bilancio statale	1	myzAnyp Aldwlp	AlmyzAnyp AlwTnyp	1
prove inconfutabili	prove evidenti	1	>dlp dAmgp	>dlp qATEp	1
spese processuali	spese dei tribunali/spese giudiziarie	2	rswm AlmHkmp		
ospedale carcerario	ospedale penitenziario	1	mst$fY Alsjn		
comportamento criminoso	comportamento illecito	1	slwk jnA}y		
effetti negativi	effetti nocivi/effetti dannosi/effetti pregiudizievoli/effetti collaterali/effetti pericolosi	5	Al|vAr AlDArp	/Al|vAr Alsy}p/Al|vAr Alslbyp/Al|vAr AljAnbyp	3
opinione dissidente			r>y mxAlf	r>yA mnfSlA	1
Stato membro	Stato contraente/Stato parte/Stato aderente	3	Aldwl Al>TrAf'	'Aldwl Al>EDA'	1
parità di remunerazione	parità salariale	1	Al>jr AlmtsAwy		
governo ad interim	governo provvisorio/governo tecnico/governo di transizione	3	Hkwmp m&qtp	Hkwmp tSryf Al>EmAl/Hkwmp AntqAlyp	2
assistenza sanitaria	assistenza medica	1	AlrEAyp AlSHyp	AlrEAyp AlTbyp	1
corte costituzionale	corte suprema	1	AlmHkmp Aldstwryp		
tutela dei diritti umani	tutela dei diritti civili	1	HmAyp Hqwq AlAnsAn	HmAyp AlHqwq Al<nsAnyp	1
uso eccessivo della forza	uso ingiustificato della forza/uso improprio della forza	2	AlAfrAT fy AstxdAm Alqwp		
centro di detenzione			mrAkz AlAHtjAz	mrkz AEtqAl	1
autorità locale	autorità municipale	1	slTp mHlyp	AlslTAt Almrkzyp	1
sistema giudiziario	sistema di giustizia	1	nZAm qDA}y	nZAm AlEdAlp	1
formazione professionale	formazione vocazionale	1	Alt>hyl Almhny	Alt>hyl wAlwZyfy	1
dispositivo di protezione individuale			mEdAt AlwqAyp Al$xSyp	mEdAt AlwqAyp Alfrdyp	1
mutilazione genitale	mutilazione sessuale	1	xtAn Al<nAv	xtAn AlftyAt	1
autorità governativa	autorità statale/autorità pubblica	2	slTp Hkwmyp	slTp Aldwlp/slTp EAmp/slTp Emwmyp	3
libertà di movimento	libertà di circolazione	1	Hryp AlHrkp	Hryp Altnql	1
guardia presidenziale			AlHrs Alr}Asy	AlHrs Aljmhwry	1
posto di lavoro			frSp Eml	frS AltwZyf	1
procedure burocratiche	procedure amministrative	1	Al<jrA'At Al<dAryp		
inchiesta ufficiale	inchiesta formale	1	tHqyq rsmy		
misure legislative			AltdAbyr Alt$ryEyp	AltdAbyr AlqAnwnyp	1
prevenzione del crimine	prevenzione del reato	1	mnE AljrA}m		
prove attendibili			>dlp mwvwq bhA	>dlp dAmgp/>dlp qATEp	2
permesso di lavoro	permesso di impiego	1	tSryH Eml		
pagamento di un'ammenda	pagamento di una multa	1	dfE grAmp		
violenza comunitaria	violenza settaria	1	AlEnf AlTA}fy	AlEnf AlmjtmEy	1
disposizioni integrative	disposizioni complementari	1	>HkAm tkmylyp		

confessioni estorte sotto tortura	confessioni estorte sotto minaccia/confessioni estorte sotto coercizione	2	AEtrAfAt mntzEp	AlAEtrAfAt bAl<krAh	1
disposizione transitoria			>HkAmAF AntqAlyp	>HkAmAF m&qtp	1
copia autenticata	copia certificata	1	nsxp mSdq ElyhA	nsxp mEtmdp	1
micro imprese	piccole imprese	1	Almn$|t AlSgyrp		
corte militare	corte marziale	1	mHkmp Eskryp		
primo ministro	-		r}ys AlwzrA'	r}ys AlHkwmp	1
matrimonio forzato			AlzwAj Alqsry	AlzwAj bAl<krAh	1
Referendum pubblico	Referendum nazionale/ referendum popolare	2	AstftA' EAm	AstftA' $Eby	1
mezzi di comunicazione elettronici	mezzi d'informazione elettronici	1	wsA}l Al<ElAm	wsA}l Al<tSAl	1
suffragio universale			AlAqtrAE AlEAm	AlAqtrAE Al$Aml	1
commissione di inchiesta	commissione investigativa	1	ljnp AltHqyq	ljnp tqSy AlHqA}q	1
pubblicazione di notizie false			n$r >xbAr kA*bp	n$r mElwmAt kA*bp/n$r >nbA' kA*bp	2
supervisione giudiziaria	supervisione di un magistrato/supervisione della magistratura	2	Al<$rAf AlqDA}y		
voto di sfiducia			AltSwyt ElY Hjb Alvqp	tSwyt bEdm Alvqp/AltSwyt bsHb Alvqp	2
tecniche di interrogatorio rinforzate	tecniche di interrogatorio abusive / tecniche di interrogatorio oltraggiose/ tecniche di interrogatorio vietate	3	>sAlyb AstjwAb m$ddp	>sAlyb AstjwAb mHrmp	1
libertà provvisoria	libertà vigilata	1	Al<frAj Alm$rwT	<frAj m&qt	1
pene detentive	pene carcerarie	1	>HkAm Alsjn		
elezioni amministrative	elezioni locali/ elezioni provinciali/ elezioni dei consigli comunali	3	AlAntxAbAt AlmHlyp	AntxAbAt bldyp	1
responsabilità penale individuale			Alms&wlyp AljnA}yp Alfrdyp	Alms&wlyp AljnA}yp Al$xSyp	1
legge sullo status personale	legge sul codice personale	1	qAnwn Al>HwAl Al$xSyp		
consiglio dei ministri	consiglio di gabinetto	1	mjls AlwzrA'		
disturbo dell'ordine pubblico	disturbo della sicurezza pubblica/disturbo alla quiete pubblica	2	tEkyr Sfw AlnZAm AlEAm		
totale	225		195		

6.3.1.3. sinonimia

Termine di base italiano	Varianti del Termine di base italiano		Termine di base arabo	Varianti del Termine di base arabo	n
funzionario pubblico	agente statale	1	mwZf Emwmy	mstxdmw Aldwlp	1
minoranze etniche	gruppi razziali	1	>qlyAt Erqyp	AlmjtmEAt Al<vnyp	1
rimpatri forzati	espulsioni irregolari/espulsioni di massa/espulsioni arbitrarie	3	AlAbEAd Alqsry	EmlyAt AltrHyl	1
difensore civico	-		AlmdAfE AlEAm	/mHAmy AlHq Almdny / >myn dywAn AlmZAlm/mHAmy AlmZAlm/ms&wl mktb AlmZAlm	4
procuratore generale	pubblico ministero/direttore della pubblica accusa/pubblica accusa	3	AlmdEy AlEAm	AlmHAmy AlEAm/mdyr AlnyAbp AlEAmp	2
agente di polizia	forze di sicurezza/autorità della sicurezza	2	DAbT $rTp	qwAt Al>mn	1
società civile	organizzazione civica	1	AlmjtmE Almdny		
procedimento giudiziario	procedure giuridiche	1	Al<jrA'At AlqDA}yp		
servizio militare			Alxdmp AlEskryp	Altjnyd AlwTny Al<lzAmy	1
discriminazione razziale			Altmyyz AlEnSry	AlkrAhyp AlErqyp	1
mandato di arresto			>mr bAlqbD	m*krp AEtqAl/m*krY twqyf	2
disposizione di legge	sensi della legislazione/norme della legislazione	2	Al>HkAm AlqAnwnyp		
polizia antisommossa			$rTp mkAfHp Al$gb	qwAt Al>mn Almrkzy	1
legislazione nazionale	legge interna	1	qAnwn wTny	AllwA}H Alqwmyp	1
emendamento alla legge	cambiamenti legislativi/modifiche al diritto	2	tEdyl qAnwn	AltgyyrAt Alt$ryEyp	1
risarcimento economico	pagamento di una ricompensa/indennizzo in denaro	2	AltEwyD AlmAlY		
giudizio di colpevolezza	verdetto di condanna/sentenza di condanna	2	Hkm Al<dAnp		
alto tradimento	lesa maestà	1	AlxyAnp AlEZmY		
polizia di rapido intervento	forza d'intervento speciale/unità di intervento speciale/battaglione d'intervento speciale	3	$rTp Altdxl AlsryE	fylq AltHrk AlsryE	1
oltraggio a pubblico ufficiale	vilipendio a pubblica autorità	1	AltEdy ElY ms&wl EAm	<hAnp mwZfyn Emwmyyn/<hAnp mwZfyn rsmyyn	2
pene detentive	condanna al carcere//condanna di reclusione/sentenza al	4	>HkAm Alsjn	Eqwbp AlHbs	1

	carcere/sanzione di reclusione				
tentato omicidio			Al$rwE fy Alqtl	mHAwlp AgtyAl	1
tratta di bambini	traffico di minore	1	AlAtjAr fy Al>TfAl		
decesso in custodia	morte in carcere	1	wfAp fy AlHjz	Almwt >vnA' AlAHtjAz	1
camera preprocessuale	dipartimento indagini/dipartimento investigativo	2	dA}rp AltHqyq	grfp mA qbl AlmHAkmp/Algrfp AlsAbqp llmHAkmp	2
detenzione pre-processuale	carcere in attesa di giudizio/arresti preventivi/custodia cautelare	3	AlAHtjAz AlsAbq llmHAkmp	AlHbs AlAHtyATy /AEtqAl AstbAqy	2
abuso d'ufficio	eccesso di potere	1	<sA'p AstxdAm AlslTp	AstglAl AlmnSb	1
immunità diplomatiche			AlHSAnAt AldblwmAsyp	Al<EfA'At AlsyAsyp	1
salute professionale	sanità del lavoro/servizi sanitari sul lavoro	2	AlSHp Almhnyp		
direzione criminale investigativa	sezione investigativa della polizia	1	<dArp AltHqyqAt AljnA}yp		
obbligo scolastico	istruzione obbligatoria	1	AltElym Al<lzAmY		
organizzazioni umanitari	enti di beneficenza	1	AlmnZmAt AltTwEyp	Alhy}At Al<nsAnyp/hy}At Al<gAvp/AlmnZmAt Al<nsAnyp/AlwkAlAt Al<nsAnyp/Alm&ssAt Alxyryp	5
perizia legale	dipartimento di medicina forense	1	AlTb Al$rEy	Altqryr AlTby-AlqAnwny/fHS Tby	2
sentenza definitiva	verdetto finale/condanna finale	2	Hkm nhA}y		
profilazione etnica			AltSwrAt AlnmTyp AlErqyp	AlAsthdAf AlEnSry	1
uguaglianza dinnanzi alla legge	eguale trattamento avanti i tribunali	1	AlmsAwAp >mAm AlqAnwn		
annullamento di un contratto			fsx Eqd	<nhA' AtfAq AlAstxdAm	1
parità delle opportunità	eguaglianza delle possibilità	1	tkAf& AlfrS		
diritto sindacale			AlHryp AlnqAbyp	Hq AltnZym	1
intercettazione di telefonate			AltnSt ElY AlAtSAlAt AlxASp	EmlyAt AlAEtrAD	1
rappresentante legale			Almmvl AlqAnwny	mHAm mwkl	1
archiviazione delle accuse			<sqAT Althm	<qfAl AlqDyp	1
rilascio condizionato	libertà provvisoria	1	Al<frAj Alm$rwT		
mandato di comparizione	istanza di habeas corpus	1	m*krAt AstdEA'	>mr bAlHDwr	1
disposizioni della sharia	leggi islamiche/norme islamiche	2	>HkAm Al$ryEp		
notifica di sgombero			<$EAr bAl<xlA'	<n*ArAF bAlmgAdrp/>mr bAl<xlA'	2
popolazioni native	gruppi di popoli indigeni/comunità indigena locale	2	AlskAn Al>Slyyn		

salute riproduttiva	assistenza sanitaria in materia di procreazione	1	AlSHp Al<njAbyp		
censura preventiva alla stampa	misure censorie contro i quotidiani	1	AlrqAbp Almsbqp ElY AlSHf		
associazione a delinquere	cospirazione criminale/impresa criminale congiunta/collusione con le bande criminali	3	Alt|mr AljnA}y	AlArtbAT mE EnASr <jrAmyp/AltwAT& lArtkAb jrymp /AltwAT& mE AlESAbAt Al<jrAmyp	3
parità di trattamento			AlmsAwAp fy AlmEAmlp	AlmEAmlp AlEAdlp	1
istruzione primaria			AltElym Al>sAsy	Sfwf AldrAsp AlAbtdA}yp	1
orientamento professionale			Altwjyh Almhny	AltElym AlwZyfy	1
sviluppo della carriera	promozione di opportunità	1	AltTwr AlwZyfy	Altrqyp Almhnyp	1
contratto di lavoro			Eqd Eml	AtfAq AlAstxdAm	1
fondi pubblici			AlSnAdyq AlEAmp	>mwAl Aldwlp/Al>mwAl AlEAmp	2
applicazione della legge	impiego della legislazione	1	tTbyq qwAnyn	<nfA* Alt$ryEAt	1
buona gestione	corretta amministrazione	1	Al<dArp Alslymp		
diniego della cittadinanza	privazione arbitraria della nazionalità/negare la nazionalità	2	Altjryd mn Aljnsyp		
brogli elettorali	manipolazione dei voti	1	tzwyr AlAntxAbAt	tlAEb fy Al>SwAt	1
persone eminenti	individui di alto profilo/figure politiche di rilievo	2	$xSyAt bArzp		
agenzie di collocamento	servizi dell'impiego	1	wkAlAt Alt$gyl		
termine ultimo	limite di tempo	1	Al>Tr Alzmnyp		
mercato del lavoro			swq AlEml	frS Almhnyp	1
misure provvisorie	provvedimenti temporanei	1	tdAbyr m&qtp		
principio del non-refoulement			mbd> HZr Al<EAdp Alqsryp	mbd> Edm AlTrd/mbd> Edm Alrd	2
quotidiano a capitale privato	giornale a proprietà privata	1	SHyfp xASp		
contabilità nazionale	conti dello stato	1	AlHsAbAt Alqwmyp		
remunerazione uguale	non discriminazione salariale/parità salariale	2	mbd> Al>jr AlmtsAwy		
aziende autogestite			Al$rkAt Alty tdAr lHsAb >SHAbhA	mn$|t mstqlp wmdArp *AtyAF	1
imprese multinazionali			Almn$|t mtEddp Aljnsyp	Al$rkAt Ebr AlwTnyp	1
approvazione parlamentare	ratifica da parte del Senato	1	AltSdyq Elyh fy mjls Al$ywx		
pianificazione familiare			tnZym Al>srp	wsA}l mnE AlHml	1
comportamento criminoso	condotta illegale	1	slwk jnA}y	mmArsAt gyr qAnwnyp	1

traffico illecito	commercio clandestino	1	AlAtjAr gyr Alm$rwE	EmlyAt thryb	1
Stato membro	paese aderente	1	Aldwl Al>TrAf		
stereotipi sessisti			AlqwAlb AlnmTyp ʿdwAr Aljnsyn	AlAnHyAz AlqA}m ElY nwE Aljns	1
azione coercitiva			tdAbyr AlqmE	<jrA' qsry/AlmmArsAt Alqsryp/AlAstxdAm Al<krAhy	3
assistenza sanitaria			AlrEAyp AlSHyp	AlxdmAt AlTbyp	1
centro di detenzione			mrAkz AlAHtjAz	tHt AltHfZ/znAzyn m$ddp AlHrAsp	2
condizioni lavorative			Zrwf AlEml	$rwT AstxdAm	1
personale di sicurezza	corpi delle forze	1	ms}wl Al>mn		
mutilazione genitale			xtAn Al<nAv	mmArsp t$wyh Al>EDA' AltnAslyp Al>nvwyp	1
autorità governativa	potere statale	1	slTp Hkwmyp	hy}p EAmp	1
regime militare			nZAm Eskry	Hkm Almjls Al>ElY llqwAt AlmslHp	1
legge interna	legislazioni nazionali	1	qAnwn dAxly		
sistema penitenziario	regime carcerario	1	nZAm Alsjwn		
pagamento di un'ammenda	contravvenzione per un reato	1	dfE grAmp		
collegio giudicante	organo della corte	1	hy}p AlmHkmp		
violazioni della privacy			AlAftqAr <lY Alsryp	xrq llxSwSyp	1
ammenda amministrativa	sanzione finanziaria	1	grAmAt AlAEtrAf bAlthmp	EqwbAt mAlyp	1
primo ministro	capo del governo/presidente del consiglio dei ministri	2	r}ys AlwzrA'		
legge sullo status personale	diritto privato	1	qAnwn Al>HwAl Al$xSyp		
detenzione arbitraria	incarceramento illegale		AlAEqAl AlEsfy		
totale	84		77		

6.3.2. Variazioni morfo-sintattiche

Termine italiano	variante	n	Termine arabo	variante	n
grazia **presidenziale**	grazia **del presidente**	1	Hqwq **Al<nsAn** (diritti umani)	Hqwqhm **Al<nsAnyp**	1
autorità **governativa**	autorità di **governo**	1	AlqAnwn **AljnA}y** (diritto penale)	qAnwn **AljnAyAt**	1
mortalità materna	**morti** materni	1	**AEtqAl** sry (detenzione segreta)	**mEtqlAt** sryp	1
ministro dell'**interno**	ministro degli **interni**	1	**Eml** Al>TfAl (lavoro minorile)	**EmAlp** Al>TfAl	1
campagna **elettorale**	campagna **per le elezioni**	1	nqAbp **EmAl** (sindacato dei lavoratori)	nqAbp **AlEAmlyn**	1
difensore civico	**difesa** civica	1	ElAqAt jnsyp gyr **$rEyp** (relazioni sessuali non legali)	ElAqAt jnsyp gyr **m$rwEp**	1
diritti **umani**	diritti **dell'uomo**	1	nZAm **qDA}y** (sistema giudiziale)	nZAm **AlqDA'**	1
disposizione **di legge**	disposizioni **legali**	1	AlmnZmAt **AltTwEyp** (organizzazioni di volontariato)	AlmnZmAt **AlTwEyp**	1
ora di **lavoro**	ora **lavorativa**	1	AlAgtSAb **Alzwjy** (stupro coniugale)	AgtSAb **Alzwjp**	1
detenzione **segreta**	detenuto **in segreto**	1	AlmwATnyn **Al>jAnb** (cittadini stranieri)	mwATnyn **>jnbyyn**	1
equo processo	**equità processuale**/**equità** del processo	2	**AlmxATr** Almhnyp (rischi professionali)	**AlAxTAr** Almhnyp	1
crimine organizzato	**criminalità** organizzata	1	Alf}At **AlmstDEfp** (comunità vulnerabili)	Alf}At **AlDEyfp**	1
modifiche c**ostituzionali**	modifiche della **costituzione**	1	**AlAnHyAz** AlsyAsy (pregiudizio politico)	**AltHyz** AlsyAsy	1
pene **detentive**	pena di **detenzione**	1	Hqwq **Alskn** (diritto all'alloggio)	Hqwq **AlmskAn**	1
ambiente **di lavoro**	ambiente **lavorativo**	1	**AlnzAE** AlmslH (conflitto armato)	**AlmnAzEAt** AlmslHp	1
sviluppi **legislativi**	sviluppi nella **legislazione**	1	**Altgyb** En AlEml (assenza dal lavoro)	**Altgyyb** En AlEml	1
giornata **di lavoro**	giornata **lavorativa**	1	AlA}tlAf **AlHAkm**(coalizione di governo)	AlA}tlAf **AlHkwmy**	1
perizia legale	**periti** legali	1	Alm$tryAt **AlEAmp** (appalti pubblici)	Alm$tryAt **AlEmwmyp**	1
sicurezza **personale**	sicurezza della **persona**	1	**AlAtfAqAt** AljmAEyp (contratti collettivi)	**>tfAq** jmAEY	1
rilascio **condizionato**	rilascio con la **condizionale**/ rilascio **condizionale**	2	<jrA'At **Altq$f** (misure d'austerità)	Al<jrA'At **Altq$fyp**	1
lacune **legislative**	lacune nella **legislazione**	1	**m$Arkp** n$Tp (partecipazione attiva)	**AlA$trAk** Aln$T	1
persone **disabili**	persone con **disabilità**	1	tlwv **Alby}p** (inquinamento ambientale)	Altlwv **Alby}y**	1
misure **organizzative**	misure di **organizzazione**	1	qAnwn **AlAntxAbAt** (legge elettorale)	qAnwn **AntxAby**	1
prevenzione del **crimine**	prevenzione di **criminalità**	1	tzwyr **AlAntxAbAt** brogli elettorali	mxAlfAt **AntxAbyp**	1

atti **di violenza**	azioni **violenti**	1	>frAd **>mnyyn** forze di sicurezza	>frAd **mn >mn**	1
garanzie **procedurali**	garanzie di **procedura**	1	**AlmdEy** AlEAm (procutatore generale)	**AlAdEA'** AlEAm	1
misure di **ispezione**	misure **ispettive**	1	**Aljrymp** AlmnZmp crimine organizzata	Aln$AT **Al<jrAmy** AlmnZm	1
mercato del **lavoro**	mercato **lavorativo**	1	tEdylAt **dstwryp** riforme costituzionali	tEdylAt **ElY Aldstwr**	1
età **lavorativa**	età di **lavoro**/Età da **lavoro**	2	**AstqlAl** AlqDA' indipendenza della magistratura	AlqDA' **Almstql**	1
organizzazione di **assistenza**	organizzazioni **assistenziale**	1	AlHrs **Alr}Asy** guardia presidenziale	>myn sr **Alr}Asp**	1
risultati **elettorali**	risultati delle **elezioni**	1	AntqA' jns Alwlyd qbl **wlAdth** Selezione sessuale prenatale	AxtyAr jns Aljnyn qbl **mwldh**	1
personale **carcerario**	personale del **carcere**	1	AlrqAbp **Almsbqp** ElY AlSHf censura preventiva alla stampa	AlrqAbp **AlsAbqp** ElY AlTbE	1
misure **correttive**	misure **di correzione**	1	tEdAd **AlskAn** AlwTny censimento nazionale	Al<HSA' **AlskAny** AlwTnyp	1
pianificazione **familiare**	pianificazione della **famiglia**	1	**AnqDA'** mdp AlHkm scadenza della condanna	**qDA'** AlEqwbp	1
statistiche del **lavoro**	statistiche **lavorative**	1	AlHq fy **AldfAE** diritto alla difesa	AlHq fY AxtyAr **mdAfE**	1
stereotipi **sessisti**	stereotipi **sessuali**	1	AlwlAyp A**lqDA}yp AlEskryp** giurisdizione militare	**AlqDA' AlEskry**	1
contaminazione **ambientale**	contaminazione **dell'ambiente**	1	Al<$rAf A**lqDA}y** supervisione giudiziaria	<$rAf **AlqDA'**	1
inchiesta **indipendente**	**indipendenza** delle indagini	1	thm **AlArhAb** reati di terrorismo	thm **ArhAbyp**	1
consiglio **di stato**	assemblea **statale**	1	ljnp **AlAntxAbAt** consiglio elettorale	ljnp **AntxAbyp**	1
reato di matrice **razziale**	reato **razzista**	1			
sicurezza **personale**	incolumità **della persona**	1			
reato **di terrorismo**	reato **terroristico**	1			
responsabilità penale **individuale**	responsabilità penale degli **individui**	1			
governo di **transito**	governo **transazionale**	1			
libertà **condizionata**	libertà **condizionale**	1			
totale	48			38	

6.3.3. variazioni sintattiche

6.3.3.1 variazioni di inserzione
- variazioni di inserzione senza modifica delle componenti del termine

Termine italiano	variante	n	Termine arabo	variante	n
violenza domestica	violenze **in ambito** domestico	1	Hkm Al<EdAm (condanna a morte)	Hkm **b**Al<EdAm (b = con)	1
funzionario pubblico	funzionari **delle amministrazioni** pubbliche	1	AlEnf fy mHyT Al>srp (violenza familiare)	AlEnf **Dd Almr>p** fy mHyT Al>srp (**Dd Almr>p = contro donna**) **>EmAl** AlEnf **AljnA}**yp fy mHyT Al>srp (**>EmAl** = atti, **AljnA} = penali**)	2
autorità competente	autorità **amministrativa** competente	1	AlAEtqAl AltEsfy arresto arbitrario	AlAEtqAl **bSwrp** tEsfyp (**bSwrp** = in modo)	1
agente di polizia	agente **del dipartimento** di polizia/agente **del corpo** di polizia	2	AlslTp AlmxtSp autorità competente	AlslTp **AlqDA}yp** AlmxtSp (AlqDA}yp = giudiziale)	1
codice penale	codice **di procedura** penale	1	m$rwE qAnwn progetto di legge	m$rwEA **l**qAnwn (l = per) /m$rwE **tEdyl** qAnwn (tEdyl = modifica)	2
servizio militare	servizio **di leva** militare	1	DAbT $rTp poliziotto	DbAT **mn** Al$rTp mn = di	1
discriminazione razziale	discriminazione **per motivi** razziali	1	Hryp AltEbyr libertà di espressione	Hryp **fy** AltEbyr fy = in	1
mandato di arresto	mandato **interstatale** di arresto	1	AlqAnwn AljnA}y diritto penale	qAnwn **AlEqwbAt** AljnA}y AlEqwbAt = pene /qAnwn **AlEdAlp** AljnA}yp AlEdAlp = giustizia	2
libertà di stampa	libertà **degli organi** di stampa	1	>mr bAlqbD mandato di arresto	>mr **mn AlmHkmp** bAlqbD mn AlmHkmp = dalla corte	1
disposizione di legge	disposizioni **contenute** nella legge/disposizione **chiave** della legge	2	qAnwn AlEfw legge di amnistia	qAnwn **b**AlEfw (b = con)	1
polizia antisommossa	polizia **in tenuta** antisommossa/polizia **in assetto** antisommossa	2	<jAzp mdfwEp Al>jr congedo pagato	Al<jAzp **Alsnwyp** AlmdfwEp Al>jr Alsnwyp = annuale	1
congedo pagato	congedo **annuale** pagato	1	AlEmAl AlmhAjryn lavoratori immigrati	EmAl **AlmnAzl** AlmhAjrwn AlmnAzl = case	1
detenzione segreta	detenzione **di sicurezza** segreta	1	Hqwq AlEmAl diritti dei lavoratori	Hqwq **jmyE** AlEmAl (jmyE = tutti)	1
status di rifugiato	status **giuridico** di rifugiato	1	qADy AltHqyq giudice inquirente	qADyA **l**ltHqyq l = per	1

diritto del lavoratore	diritti **umani** dei lavoratori	1	qAnwn wTny diritto nazionale	AlqAnwn **AljnA}y** AlwTny AljnA}y = penale	1
giudice inquirente	giudice **co-**inquirente	1	Aljrymp AlmnZmp crimine organizzato	Aljrymp **Aldwlyp** AlmnZmp Aldwlyp = internazionale	1
incapacità delle autorità	incapacità **da parte** delle autorità	1	tEdyl qAnwn emendamento di legge	tEdyl **ElY** AlqAnwn (ElY = su) /tEdylAt **l**qAnwn (l = per) /tEdyl **>HkAm** AlqAnwn (>HkAm = sensi)	3
matrimonio omosessuale	matrimonio **tra persone** omosessuali	1	AlmHkmp Al<dAryp corte amministrativa	mHkmp **AlqDA'** Al<dAry AlqDA' = giuridizione	1
crimine organizzato	crimine **transnazionale** organizzato	1	AxtSAS AlmHkmp competenza della corte	AxtSAS **l**lmHkmp l = per	1
emendamento alla legge	emendamenti **provvisori** alla legge/emendamenti **preliminari** alla legge	2	Hkm Al<dAnp sentenza di condanna	Hkm **b**Al<dAnp b = con	1
tribunale civile	tribunale **ordinario** civile	1	Hkm mHkmp sentenza della corte	AlHkm **AlSAdr En** mHkmp AlSAdr En = emesso da	1
risarcimento economico	risarcimento **di natura** economica	1	gsyl >mwAl riciclaggio di denaro	gsyl **l**l>mwAl l = per	1
sindacato dei lavoratori	sindacato **indipendente** dei lavoratori/sindacato **nazionale** dei lavoratori	2	>HkAm Alsjn sentenze di carcere	>HkAm **b**Alsjn b = con	1
corte amministrativa	corte **suprema** amministrativa/corte **della magistratura** amministrativa	2	AlqAnwn Aldwly diritto internazionale	qAnwn **AljrA}m** Aldwlyp (AljrA}m = crimini) /AlqAnwn **Al<nsAny** Aldwly (Al<nsAny = umanitario) /AlqAnwn **AlErfy** Aldwly (AlErfy = consuetudinale)	3
traffico di droga	traffico **illecito** di droga/reati **in materia** di droga	2	<dArp AltHqyqAt AljnA}yp Direzione criminale investigativa	Al<dArp **AlwTnyp l** AltHqyqAt AljnA}yp AlwTnyp l = nazionale per	1
codice di condotta	codici **volontari** di condotta/codice **internazionale** di condotta	2	AlHq fy Alg*A' diritto al cibo	AlHq fy **AlHSwl ElY** Alg*A' AlHSwl ElY = ottenere	1
appropriazione di fondi	appropriazione **indebita** di fondi/appropriazione i**ndebita di denaro** del fondo	2	AlAntqAl AldymqrATy transizione democratica	AlAntqAl **<lY AlHkm** AldymqrATy <lY AlHkm = verso il governo	1
processo di corruzione	processo **per accuse** di corruzione	1	AlmmArsAt Altmyyzyp trattamento discriminatorio	AlmmArsAt **AlAjtmAEyp** Altmyyzyp AlAjtmAEyp = sociale	1
ministro dell'interno	ministro degli **affari** interni	1	AlmsAwAp >mAm AlqAnwn eguaglianza innanzi alla legge	AlmsAwAp **AltAmp lljmyE** >mAm AlqAnwn AltAmp lljmyE = tutta per tutti	1
immunità giudiziaria	immunità **dai procedimenti** giudiziari/immunità **dall'azione** giudiziario/immunità **dal perseguimento** giudiziario	3	Altnmyp AlmstdAmp sviluppo sostenibile	Altnmyp **Alb$ryp** AlmstdAmp Alb$ryp = umana	1
sviluppi legislativi	sviluppi **in ambito** legislativo	1	AlHryp AlnqAbyp libertà sindacale	Hryp **mmArsp AlEml** AlnqAby mmArsp AlEml = prassi	1

inchiesta indipendente	inchiesta **internazionale** indipendente	1	txfyf llEqwbp riduzione della pena	txfyf **Hkm** AlEqwbp Hkm = sentenza	1
perizia legale	perizia **medico-** legale	1	Alt>myn AlAjtmAEY assicurazione sociale	Alt>myn AlAjtmAEY **Al<lzAmY** Al<lzAmY = obbligatorio	1
sentenza definitiva	sentenza **in via** definitiva	1	AlmxATr Almhnyp rischi professionali	AlmxATr **AlSHyp** Almhnyp (AlSHyp = di salute) /mxATr **AltErD** Almhny (AltErD = esposizione)	2
testimone dell'accusa	testimoni della **pubblica** accusa/testimone **chiave** della accusa	2	jrymp $rf delitto d'onore	jrA}m **bdAfE** Al$rf bdAfE = per motivi	1
sicurezza nazionale	sicurezza **sociale** nazionale	1	Alf}At AlmstDEfp comunità vulnerabili	Alf}At **AlskAnyp** AlmstDEfp AlskAnyp = di popolazione	1
diritto alla difesa	diritti **dei detenuti** alla difesa	1	AlqwAnyn AlErfyp leggi consuetudinarie	AlqAnwn **Aldwly** AlErfy Aldwly = internazionale	1
camera d'appello	camera **penale della corte** d'appello	1	Hqwq Alskn diritti all'alloggio	AlHq **fy** Alskn fy = in	1
prassi discriminatorie	prassi **culturali** discriminatorie	1	AlmxATr AlSHyp rischi alla salute	mxATr **Al>DrAr** AlSHyp Al>DrAr = danni	1
camera preliminare	camera **dei giudizi** preliminari	1	ljnp mnAhDp AltE*yb AltAbEp ll>mm AlmtHdp Comitato delle Nazioni Unite contro la tortura	Alljnp **AlmEnyp b**mnAhDp AltE*yb AltAbEp ll>mm AlmtHdp AlmEnyp b= interessata a	1
proventi del reato	proventi **relativi** ai reati	1	AlnzAE AlmslH conflitto armato	AlnzAE **AldAxly** AlmslH AldAxly = interno	1
soluzione delle controversie	soluzione **pacifica** delle controversie	1	Al<qAmp Aljbryp arresto domiciliare	Al<qAmp **Almnzlyp** Aljbryp mnAzlhm = in casa	1
rischi professionali	rischi di **esposizione** professionale	1	tgyb Al$hwd assenteismo dei testimoni	tgyb Al$hwd **En AlHDwr** En AlHDwr = assistere	1
oltraggio alla corte	oltraggio **criminale** alla corte	1	AltElym Al>sAsy insegnamento formale	AltElym **AlmjAny** Al>sAsy AlmjAny = gratuito	1
giurisdizione militare	giurisdizione **dei tribunali** militari	1	<nhA' AlAtfAq cessazione del contratto	Al<nhA' **Almbkr l**AtfAq Almbkr l= anticipato di	1
delitto d'onore	delitto **per motivi** d'onore	1	tTbyq qwAnyn applicazione delle leggi	tTbyq **l>HkAm** qAnwn l>HkAm = disposizioni	1
gruppi vulnerabili	gruppi **particolarmente** vulnerabili	1	>TrAf AlnzAE parti in conflitto	Al>TrAf **fy** AlnzAE fy = in	1
cospirazione criminale	cospirazione **con finalità** criminali	1	AltElym Alrsmy istruzione formale	**brAmj** AltElym Alrsmy brAmj = programmi	1
rilascio condizionato	rilasciato **in libertà** condizionata	1	Hq AlEml diritto al lavoro	AlHq **fy** AlEml fy = in	1
rilascio anticipato	rilascio **condizionale** anticipato	1	Al<dArp Alslymp buona gestione	Al<dArp **Alby}yp** Alslymp Alby}yp = ambientale	1
forze filogovernative	forze **di sicurezza**	1	r}ys AlwzrA'	r}ys **mjls** AlwzrA'	1

	filogovernative		primo ministro	mjls = consiglio	
sentenza della corte	sentenza **emessa** dalla corte	1	AstftA' tqryr AlmSyr referendum sull'autodeterminazione	AstftA' **b$>n** tqryr AlmSyr (b$>n = riguardo) /AstftA' **ElY** tqryr AlmSyr/ (ElY = su) AstftA' **En** tqryr mSyr (En = di)	3
diritti fondamentali sul lavoro	diritti fondamentali sul **luogo** di lavoro	1	AsthdAf Erqy profilazione etnica	AsthdAf **>bnA' AljmAEAt** AlErqyp >bnA' AljmAEAt = figli delle comunità	1
rischi per la salute	rischi **esistenti** per la salute	1	wzyr AldAxlyp ministro dell'interno	wzyr **Al$&wn** AldAxlyp Al$&wn = affari	1
conflitto armato	conflitto **interno** armato	1	qwAt HfZ AlslAm forze di peacekeaping	qwp **l**HfZ AlslAm l = per	1
violenza comunitaria	violenza **inter-**comunitaria	1	tEdAd AlskAn AlwTny censimento nazionale	AltEdAd AlskAny **ElY AlSEyd** AlwTnyy ElY AlSEyd = su livello	1
assenza dal lavoro	assenze **giustificate** dal lavoro	1	qtl syAsy omicidio politico	qtl ***At TAbE** syAsy (*At TabE = da qualità) /Alqtl **ldwAfE** syAsyp (ldwAfE = per motivi) /'qtl **Aln$TA'** AlsyAsyyn (Aln$TA' = attivisti) /qtl **b dwAfE** syAsyp (con motivi)	4
diritto alla riservatezza	diritto **al rispetto** della riservatezza	1	zEzEp AlHkm destabilizzare il regime	zEzEp **AstqrAr nZAm** AlHkm AstqrAr nZAm = stabilità del regime	1
condizioni di impiego	condizioni **determinate** di impiego/condizioni **minime** di impiego	2	tzwyr AlAntxAbAt broglio elettorale	tzwyr **fy** AlAntxAbAt fy = in	1
cessazione del contratto	cessazione **anticipata** del contratto	1	Hq Almlkyp diritto alla proprietà	AlHq **fy** Almlkyp fy = in	1
applicazione della legge	applicazione **extraterritoriale** delle leggi/applicazione **concreta** delle leggi	2	AljmEyp AlwTnyp associazione nazionale	jmEyp **AlEml** AlwTny (AlEml = lavoro) /AljmEyp **Alt$ryEyp** AlwTnyp (Alt$ryEyp = legislativo)	2
parti del conflitto	parti **coinvolte** del conflitto	1	tDArb AlmSAlH conflitto di interess	tDArb **fy** AlmSAlH fy = in	1
misure organizzative	misure **di natura** organizzativa	1	nZAm tfty$ sistema di ispezione	nZAm **l**ltfty$ l = per	1
accuse di corruzione	accuse **pretestuose** di corruzione	1	AlAtfAqAt AljmAEyp contrattazione collettiva	AtfAq **AlmfAwDp** AljmAEyp AlmfAwDp = commissione	1
omicidi politici	omicidi **di matrice** politica	1	<EAdp AllAj}yn Rientro di profughi	<EAdp **twTyn** lAj}yn twTyn = insediamento	1
mancanza di sicurezza	mancanza **di condizioni** di sicurezza	1	lm $ml >srp ricongiungimento familiare	lm $ml **>frAd** Asrp >frAd = membri	1
bozza di costituzione	bozza **di legge** costituzionale	1	sn AlEml età lavorativa	sn >dnY llAltHAq bAlEml >dnY llAltHAq = minimo per svolgere	1
diritto di proprietà	diritto di **possedere** proprietà	1	SHyfp xASp quotidiano privato	AlSHf ***At Almlkyp** AlxASp *At Almlkyp = con proprietà	1
assemblea nazionale	assemblea **legislativa** nazionale	1	>mr AlmSAdrp ordine di confisca	>mr **b**mSAdrp b = con	1

mercato del lavoro	mercato **libero** del lavoro/mercato **aperto** del lavoro	2	tbAdl AlmElwmAt scambio di informazioni	tbAdl **mntZm** llmElwmAt mntZm = regolare	1
misure provvisorie	misure **a titolo** provvisorio	1	mnE AljrA}m prevenzione del crimine	mnE **wqwE** Aljrymp wqwE = avvenire	1
scambio di informazioni	scambio **sistematico** di informazioni	1	sjl jnA}y casellario penale	sjl **AlswAbq** AljnA}yp AlswAbq = precedenti	1
servizi pubblici	servizi **di utilità** pubblica	1	ms&wly Alsjwn personale di carcere	ms&wlwn **En** Alsjwn En = di	1
periodi di prescrizione	periodo **prolungato** di prescrizione	1	jmAEp mEArDp gruppo di opposizione	jmAEAt **dynyp** mEArDp (dynyp = religiosa) /AljmAEAt **AlsyAsyp** AlmEArDp (AlsyAsyp = politico)	2
salario base	salario **di** base/salario **minimo** di base	1	AlHkwmp Alm&qtp governo ad interim	Hkwmp **ltSryf Al>EmAl** m&qtp ltSryf Al>EmAl = provvisoriamente	1
disarmo civile	disarmo **dei** civili	1	mnZmp Hqwq Al<nsAn organizzazione dei diritti umani	mnZmAt **mEnyp** bHqwq Al<nsAn (mEnyp = interessata) /mnZmAt **mHlyp mEnyp** bHqwq Al<nsAn (mHlyp mEnyp = locali e interessati)	2
permesso di soggiorno	permessi **temporanei** di soggiorno	1	AltdAbyr Alt$ryEyp misure legislative	AltdAbyr **AlmlA}mp fy AlmjAlAt** Alt$ryEyp AlmlA}mp fy AlmjAlAt = convenienti nei campi	1
organizzazione di assistenza	organizzazione **dei servizi** di assistenza	1	tnfy* AltwSyAt applicare le raccomandazioni	Altnfy* **AlfEAl** ltwSyAt AlfEAl = effettivo	1
bilancio nazionale	bilancio **a livello** nazionale	1	ljnp Hqwq Al<nsAn commissione dei diritti umani	ljnp **wTnyp wmstqlp** lHqwq Al<nsAn (wTnyp wmstqlp = nazionale indipendente) / Alljnp **AlfrEyp l**Hqwq Al<nsAn AlfrEyp = sotto	2
personale carcerario	personale **del servizio** carcerario	1	mjls AlAntxAbAt consiglio elettorale	Almjls **Al>ElY** llAntxAbAt Al>ElY = superiore	1
opinione dissidente	opinioni **politiche** dissidenti	1	thm AlArhAb	thm t**tElq b**AlArhAb/thmp ***At Slp b**Al<rhAb ttElq b,*At Slp b = riguardante	2
norme di contabilità	norme **in materia** di contabilità	1	AlAntxAbAt AlmHlyp elezioni amministrative	AntxAbAt **AlmjAls** AlmHlyp/ AntxAbAt **AlHkwmAt** AlmHlyp AlmjAls = consiglio AlHkwmAt = governi	2
statistiche del lavoro	statistiche **di base** del lavoro	1	AnthAkAt Hqwq Al<nsAn	AnthAkAt **jsymp** lHqwq Al<nsAn jsymp = gravi	1
organizzazioni dei diritti umani	organizzazioni **di difesa** dei diritti umani/organizzazioni **di monitoraggio** del diritto umano	2	Hkwmp tSryf Al>EmAl governo di transito	Hkwmp **l**tSryf Al>EmAl l = per	1
misure legislative	misure **in campo** legislativo/misure **di ordine** legislativo	2	tjArp AlmxdrAt traffico di droga	AlAtjAr **b**AlmxdrAt b = con	1
ora di lavoro	ore **massime** di lavoro/ore	2	wvA}q Alhwyp	wvA}q **<vbAt** Alhwyp	1

	normali di lavoro		documenti di identità	<vbAt = attestare	
permesso di lavoro	permesso **temporaneo** di lavoro	1	gsyl Al>mwAl riciclaggio del denaro	gsyl **l**l>mwAl l = di	1
gruppi religiosi	gruppi **di fede** religiosa/gruppi **di opposizione** religiosa	2			
protezione dei civili	protezione delle **popolazioni** civili	1			
gruppi di opposizione	gruppi **politici** di opposizione/gruppo di opposizione **armata**/gruppi **armati** di opposizione	3			
misure legislative	misure **in campo** legislativo/misure **di ordine** legislativo	2			
reato di terrorismo	reati collegati al terrorismo	1			
totale	114		106		

- variazioni di inserzione con modifica delle componenti del termine

Termine di base italiano	Varianti del Termine di base italiano	n	Termine di base arabo	Varianti del Termine di base arabo	n
diritti umani	diritti **fondamentali della persona**	1	m$rwE qAnwn progetto di legge	**mswdp jdydp l**qAnwn progetto nuovo di legge	1
violenze sessuali	**abuso di natura** sessuale	1	Altmyyz AlEnSry discriminazione razziale	Altmyyz **AlmbA$r wgyr AlmbA$r ElY >sAs Aljns** discriminazione razziale diretta e indiretta	1
disposizione di legge	disposizioni **applicabili della legislazione**	1	mnZmAt AlmjtmE Almdny organizzazioni della società civile	**jmAEAt mn** AlmjtmE Almdny gruppi della società civile	1
servizio militare	**arruolamento** militare **nazionale**	1	AlHq fy Alg*A' diritto al cibo	AlHq fy **AlHSwl ElY AlTEAm** diritto a procurarsi il cibo	1
contrattazione collettiva	**negoziare a livello** collettivo	1	<jAzp AlwDE congedo di maternità	<jAzp **<lzAmyp bEd AlwlAdp** congedo di maternità obbligatoria	1
detenzione amministrativa	**arresti per reati** amministrativi	1	AlAnHyAz AlErqy pregiudizi etnici	AnHyAzhA A**lqA}m ElY >sAs AlAntmA' Aljnsy** pregiudizi basati sul sesso	1
equo processo	equo **procedimento giudiziario**	1	AlAstnkAf AlDmyry En Alxdmp AlEskryp obiezione di coscienza	**AlAEtrAD ElY t>dyp** Alxdmp AlEskryp **bdAfE** AlDmyr obiezione di coscienza	1
incapacità delle autorità	**refoulement da parte** delle	1	AzdrA' AlmHkmp	**AlAHtqAr AljnA}y** llmHkmp	2

	autorità		oltraggio alla corte	oltraggio penale alla corte AnthAk Hrmp AlmHkmp oltraggio alla corte	
matrimonio omosessuale	**relazione amorosa** omosessuale	1	<$EAr bAl<xlA' notifica di sgombero	**<n*ArAF msbqAF** bAl<xlA' preavviso di sgombero	1
codice di condotta	**regole fondamentali** di condotta	1	Alt|mr AljnA}y associazione a delinquere	**AltwrT fy qDAyA** jnA}yp coinvolgimento in cause penali	1
cittadini stranieri	**persone di nazionalità** straniera	1	Altjryd mn Aljnsyp diniego della cittadinanza	**AlHrmAn AltEsfy** mn Aljnsyp diniego arbitrario della cittadinanza	1
inchiesta indipendente	**commissione** indipendente **sulle indagini**	1	>HkAm mE wqf Altnfy* sentenze con sospensione della pena	>HkAm **gyr** nAf*p pena sospesa	1
pregiudizi etnici	pregiudizio **basato sull'identità sessuale**	1	$xSyAt bArzp persone eminenti	$xSyAt **syAsyp rfyEp AlmstwY** persone politiche eminenti	1
riduzione della pena	**attenuazione** della pena **prevista**	2	<tlAf AlmmtlkAt AlEAmp distruzione di proprietà pubblica	**<lHAq AlDrr b$kl mtEmd** bAlmmtlkAt AlEAmp distruzione premeditata di proprietà pubblica	1
prove attendibili	**elementi** di prova **rilevanti**	1	mEdl AlbTAlp tasso di disoccupazione	**AlmstwY Alrsmy** llbTAlp tasso di disoccupazione ufficiale	1
gruppi vulnerabili	**persone più** vulnerabili/**comunità più** vulnerabili	2	sn AlEml età lavorativa	**AlHd Al>dnY** lsn **Al>stxdAm** età minima di lavoro	1
riciclaggio di denaro	riciclaggio di **fondi pubblici**	1	Alt>myn Al<lzAmy assicurazione obbligatoria	**nZAm** llt>myn **AlAjtmAEY AlAjbArY** assicurazione sociale obbligatoria	1
associazione a delinquere	**cospirazione finalizzata** a delinquere	1	qAnwn AlAntxAbAt codice elettorale	**AlqwAEd AlxASp b**AlAntxAbAt codice riguardante le elezioni	1
commissione parlamentare sui diritti umani	**comitato** parlamentare **congiunto** sui diritti umani	1	AlrEAyp AlSHyp Assistenza sanitaria	**AlrEAyp** AlTbyp **AlmlA}mp** Assistenza sanitaria appropriata /**AlmsAEdp** AlTbyp **AlkAfyp** Assistenza sanitaria sufficiente	2
false accuse	accuse **di reato costruite**/accuse **penali pretestuose**	2	HmAyp Hqwq AlAnsAn tutela dei diritti umani	**dfAE En** Hqwq Al<nsAn difesa dei diritti umani	1
permesso di soggiorno	permessi **temporanei di residenza**	1	msyrp slmyp Manifestazione pacifica	**AlmZAhrAt AlAHtjAjyp** Alslmyp Manifestazione pacifica	1
contrattazione collettiva	negoziare **a livello collettivo**	1	mHArbp AlEnf Lotta alla violenza	**lltglb ElY >EmAl** AlEnf Lotta ai atti di violenza	1
misure di ispezione	**sistema efficace** di ispezione/**modalità generali** di ispezione	2	HAlp tE*yb segnalazioni di tortura	**AlAdEA'At AlmtElqp** bAltE*yb segnalazioni riguardanti la tortura	1
metodi di pagamento	**mezzo legale** di pagamento	1	Alxdmp AlEskryp servizio militare	Alxdmp **AlwTnyp Al<lzAmyp** servizio militare obbligatorio	1
sciopero generale	sciopero **a livello nazionale**	1	Altmyyz AlEnSry discriminazione razziale	Altmyyz **Dd AlskAn Al>Slyyn** discriminazione razziale contro la	1

				popolazione nativa	
periodo di preavviso	**durata minima** di preavviso	1	AEtqAl sry detenzione segreto	**mwAqE AHtjAz** sryp centri di detenzione segreti	1
			kbt AlmEArDp opprimere l'opposizione	**qDA' ElY** AlmEArDp **AlsyAsyp** opprimere l'opposizione politica	1
			mjls AlAntxAbAt consiglio elettorale	**Alhy}p AlwTnyp l**lAntxAbAt ente nazionale per le elezioni /**Alljnp Almstqlp l**lAntxAbAt commissione indipendente per le elezioni	2
totale	30		31		

6.3.3.2 variazioni di coordinazione

Termine italiano	variante	n	Termine arabo	variante	n
violenza domestica	violenza **sessuale** e domestica	1	>qlyAt Erqyp minoranza etnica	Al>qlyAt **AlEnSryp** wAlErqyp (AlEnSryp = razziale) /Al>qlyAt **A ̓vnyp** wAlErqyp (A ̓vnyp = di razza)	2
minoranze etniche	minoranze **razziali** ed etniche	1	zwAj mvly matrimonio omosessuale	zwAj **>w AtHAd** mvly >w AtHAd = unione	1
detenzione arbitraria	detenzione **illegale** o arbitraria	1	AlHSAnAt Aldblwm Asyp immunità diplomatiche	AlHSAnAt w**AltshylAt** Aldblwm Asyp AltshylAt = privilegi	1
diritti umani	diritti **civili** e umani/diritto **umanitario** e umano	2	AlSHp Almhnyp salute professionale	AlSHp **wAlslAmp** Almhnyp AlslAmp = sicurezza	1
violenze sessuali	violenza **di genere** e sessuale	1	HmAyp Al$hwd protezione dei testimoni	**dEm** Al$hwd wHmAythm dEm = sostegno	1
libertà di espressione	libertà **di opinione** ed espressione	1	Alt>hyl Almhny riabilitazione professionale	Alt>hyl Almhny w**AlwZyfy** AlwZyfy = di carriera	1
equo processo	processo **tempestivo** ed equo/equo e **regolare** processo/processo equo e **imparziale**	3	$rwT >stxdAm Condizioni di impiego	$rwT w**>HkAm** AstxdAm >HkAm = regole	1
matrimonio omosessuale	matrimonio o **unione** omosessuale	1	tTbyq qwAnyn applicazione della legge	tTbyq AlqwAnyn w**AllwA}H** AllwA}H = regolamenti	1
misure disciplinari	misure **giudiziarie** o disciplinari	1	AlDmAnAt Al<jrA}yp garanzie procedurali	AlDmAnAt **AlqAnwnyp** wAl<jrA}yp AlqAnwnyp = giudiziale	1
immunità diplomatiche	immunità e **agevolazioni** diplomatiche	1	Alt>hyl Almhny Formazione professionale	**Altdryb** wAlt>hy**l** Almhny Altdryb = addestramento	1
inchiesta indipendente	inchiesta **approfondita** e	2	AltdAbyr Alt$ryEyp	AltdAbyr Alt$ryEyp >w	1

	indipendente/inchiesta indipendente e **imparziale**			Misure legislative	**AlqDA}yp** AlqDA}yp = giudiziale	
mandato di comparizione	mandato **d'arresto** o di comparizione	1		n$r >xbAr kA*bp pubblicazione di notizie false	n$r >xbAr w**byAnAt** kA*bp byAnAt = informazioni	1
lavoro dignitoso	lavoro **durevole** e dignitoso	1		jmwd syAsy stallo politica	Aljmwd >w **Al$ll** AlsyAsy Al$ll = paralisi	1
salute riproduttiva	salute **sessuale** e riproduttiva	1		<$rAf qDA}y supervisione giudiziaria	<$rAf >w **rqAbp** qDA}yp rqAbp = controllo	1
garanzie procedurali	tutele **legislative** e procedurali	1		Altjnyd AlqsrY reclutamento forzato	Altjnyd AlqsrY >w **AlAjbArY** AlAjbArY = obbligatorio	1
assemblea nazionale	assemblea **statale** e nazionale	1		AlAxtfA' Alqsry sparizione forzata	AlAxtfA'At **gyr AlTwEyp** wAlqsryp gyr AlTwEyp = involontario	1
misure di ispezione	misure di **controllo** e di ispezione	1				
aziende autogestite	imprese **autonome** e autogestite	1				
crisi economica	crisi **finanziaria** ed economica	1				
supervisione giudiziaria	supervisione o alcun **controllo** giudiziario					
uso eccessivo della forza	uso eccessivo e **indebito** della forza/uso eccessivo e **letale** della forza	2				
totale	25				17	

- variazioni di coordinazione con modifica delle componenti del termine

Termine di base italiano	Varianti del Termine di base italiano	n		Termine di base arabo	Varianti del Termine di base arabo	n
minoranze etniche	gruppi razziali ed etnici	1		AlAEtqAl AltEsfy detenzione arbitraria	twqyf >w AEtqAl gyr qAnwny arresto e detenzione arbitrari	1
procedimento giudiziario	procedure amministrative e legali	1		AltHryD ElY Alftnp AlTA}fyp incitamento alla lotta settaria	AvArp AlnErAt AlTA}fyp wAlm*hbyp incitamento alla lotta settaria e confessionale	1
mandato di arresto	richiesta di fermo o di arresto	1		AlSHp Almhnyp salute professionale	SHp wslAmp AlEmAl salute e sicurezza dei lavoratori	1
inchiesta indipendente	indagini credibili e indipendenti	1		thmp Alt$hyr accusa di diffamazione	thmp Alsb wAlq*f accusa di diffamazione e di oltraggio	1
rischi professionali	rischi o i pericoli riconducibili al	1		AlEnf AlTA}fy	AlEnf AlmjtmEy wAlErqy	1

	lavoro			Violenza comunitaria	Violenza comunitaria e etnica	
rilascio anticipato	proscioglimento anticipato o condizionale	1				
tecniche di interrogatorio	metodi e pratiche d'interrogatorio	1				
mercato del lavoro	politiche del lavoro e dell'occupazione	1				
distruzione di proprietà pubblica	distruggere o danneggiare beni pubblici	1				
segnalazioni di tortura	casi di tortura e maltrattamenti	1				
procedure burocratiche	procedure amministrative e legali	1				
prove attendibili	prove affidabili e sostanziali	1				
pubblicazione di notizie false	diffusione di informazioni false e esagerate	1				
molestia sessuale	vessazione verbale e sessuale	1				
oltraggio a pubblico ufficiale	blasfemia e oltraggio a pubblico ufficiale/violenza e oltraggio a pubblico ufficiale	2				
totale	16			5		

6.3.3.3 variazioni di permutazione

Termine italiano	variante	n		Termine arabo	variante	n
autorità competente	competente autorità	1		mHAkmp EAdlp equo processo	EdAlp AlmHAkmAt	1
polizia di rapido intervento	polizia di intervento rapido	1		mEArD syAsy dissidente politico	AlsyAsywn AlmEArDwn	1
selezione sessuale prenatale	selezione prenatale del sesso	1		Almjls Al>ElY llqDA' consiglio supremo della magistratura	mjls AlqDA' Al>ElY	1
possesso illegale di armi	possesso di armi illegale	1		AstqlAl AlqDA' indipendenza della magistratura	AlqDA' Almstql	1
coalizione di governo	governo di coalizione	1		AlmsAwAp fy AlmEAmlp parità di trattamento	AlmEAmlp AlmtsAwyp	1
diritto internazionale umanitario	diritto umanitario internazionale	1		AlA}tlAf AlHAkm coalizione di governo	Hkwmp A}tlAfyp	1

uso sproporzionato della forza	uso di forza sproporzionata	1		>ElY AlSfAt Alxlqyp alta autorità morale	AlSfAt Alxlqyp AlEAlyp	1
consiglio supremo della magistratura	alto consiglio della magistratura	1		AlAfrAT fy AstxdAm Alqwp Uso eccessivo della forza	AlAstxdAm AlmfrT llqwp/AstxdAm Alqwp AlmfrTp	2
direzione criminale investigativa	unità investigativa criminale/polizia investigativa criminale	2		>HkAm mE wqf Altnfy* sentenze con sospensione della pena	wqf tnfy* >HkAm	1
dispositivo di protezione individuale	dispositivi personali di protezione	1		AlslTp AlmxtSp autorità competente	>xtSAS AlslTp	1
coinvolgimento attivo	attiva partecipazione	1		>srY AlHrb prigionieri di guerra	Aljnwd Al>srY	1
ospedale carcerario	reparto penitenziario di un ospedale	1		Alljnp AlbrlmAnyp lHqwq Al<nsAn Commissione parlamentare sui diritti umani	ljnp Hqwq Al<nsAn fy mjls Al>mp /Alljnp AlmEnyp bHqwq Al<nsAn fy mjls AlnwAb	2
parità di remunerazione	remunerazione equa/remunerazione uguale	2		Al<dArp Alslymp buona gestione	Hsn AdArp	1
				>ElY AlSfAt Alxlqyp alta levatura morale	AlmnAqb Alxlqyp AlrfyEp	1
totale	15			16		

6.3.3.4 variazioni di sostituzione di preposizione

Termine italiano	variante	n		Termine arabo	variante	n
congedo **di** maternità	congedo **per** maternità	1		Al$rwE **fy** Alqtl tentato omicidio	$rwE **b**Alqtl	1
testimoni **della** difesa	testimoni **a** difesa	1		AlAtjAr **fy** Al>TfAl tratta di bambini	Al<tjAr **b**Al>TfAl	1
uguaglianza **dinnanzi** alla legge	uguaglianza **di fronte** alla legge	1		wfAp **fy** AlHjz decesso in custodia	AlwfyAt **>vnA'** AlAHtjAz	1
diritti fondamentali **sul** lavoro	diritti fondamentali **nel** lavoro	1		HSAnpF **mn** AlmqADAp immunità giudiziaria	AlHSAnp **Dd** AlmqADAp	1
parti **del** conflitto	parti **in** conflitto/parti **al** conflitto	2		AlmsAwAp **>mAm** AlqAnwn uguaglianza dinnanzi alla legge	AlmsAwAp **fy** AlqAnwn	1
scandalo **di** corruzione	scandalo **per** corruzione	1		AlHqwq Al>sAsyp **fY** AlEml diritti fondamentali sul lavoro	AlHqwq Al>sAsyp **l**lEmAl	1

organizzazioni **dei** diritti umani	organizzazioni **per** i diritti umani	1				
lotta alla violenza	lotta **contro** la violenza	1				
commissione **dei** diritti umani	commissione **per** i diritti umani/commissione **sui** diritti umani	2				
decesso **in** custodia	decessi **durante** la custodia	1				
diritto **alla** difesa	diritto **della** difesa	1				
legislazione **anti**-terrorismo	legge **contro** il terrorismo	1				
totale	14			6		

6.3.3.5 variazioni di omissione

Termine italiano	variante	n		Termine arabo	variante	n
oltraggio a **pubblica** autorità	oltraggio all'autorità	1		Hkwmp wHdp **wTnyp** governo di unità nazionale	Hkwmp wHdp	1
consiglio **supremo** della magistratura	consiglio della magistratura	1		AlrqAbp **Almsbqp** ElY AlSHf censura preventiva alla stampa	AlrqAbp ElY AlSHf	1
sentenze con sospensione della pena	sospensione della pena	1		$rTp **mkAfHp** Al$gb polizia antisommossa	$rTp Al$gb	1
rapina **a mano** armata	rapina armata	1		mnZmAt **AlmjtmE** Almdny organizzazioni della società civile	AlmnZmAt Almdnyp	1
sistema **di giustizia** militare	giustizia militare	1		AltHryD ElY **Alftnp** AlTA}fyp incitamento alla lotta settaria	AltHryD ElY AlTA}fyp	1
boicottaggio **delle elezioni** parlamentari	boicottaggio parlamentare	1		AlHq **fy** Alg*A' diritto al cibo	Hqwq Alg*A'	1
servizio militare **nazionale**	servizio militare	1		AlHq **fy** AldfAE diritto alla difesa	Hq AldfAE	1
				txfyf **l**lEqwbp riduzione di pena	txfyD AlEqwbp	1
				Alxdmp AlEskryp **AlAzAmyp** servizio militare obbligatorio	Alxdmp AlEskryp	1
totale	7			9		

6.3.5- variazioni ortografiche

Termine di base italiano	Varianti del Termine di base italiano	n
legislazione anti-terrorismo	legislazione antiterrorismo	1
stato di emergenza	stato d’emergenza	1
micro imprese	micro-imprese	1
polizia antisommossa	polizia anti-sommossa	1
diritto all'auto-determinazione	diritto all'autodeterminazione	1
periodo postelettorale	periodo post elettorale	1
Totale	6	

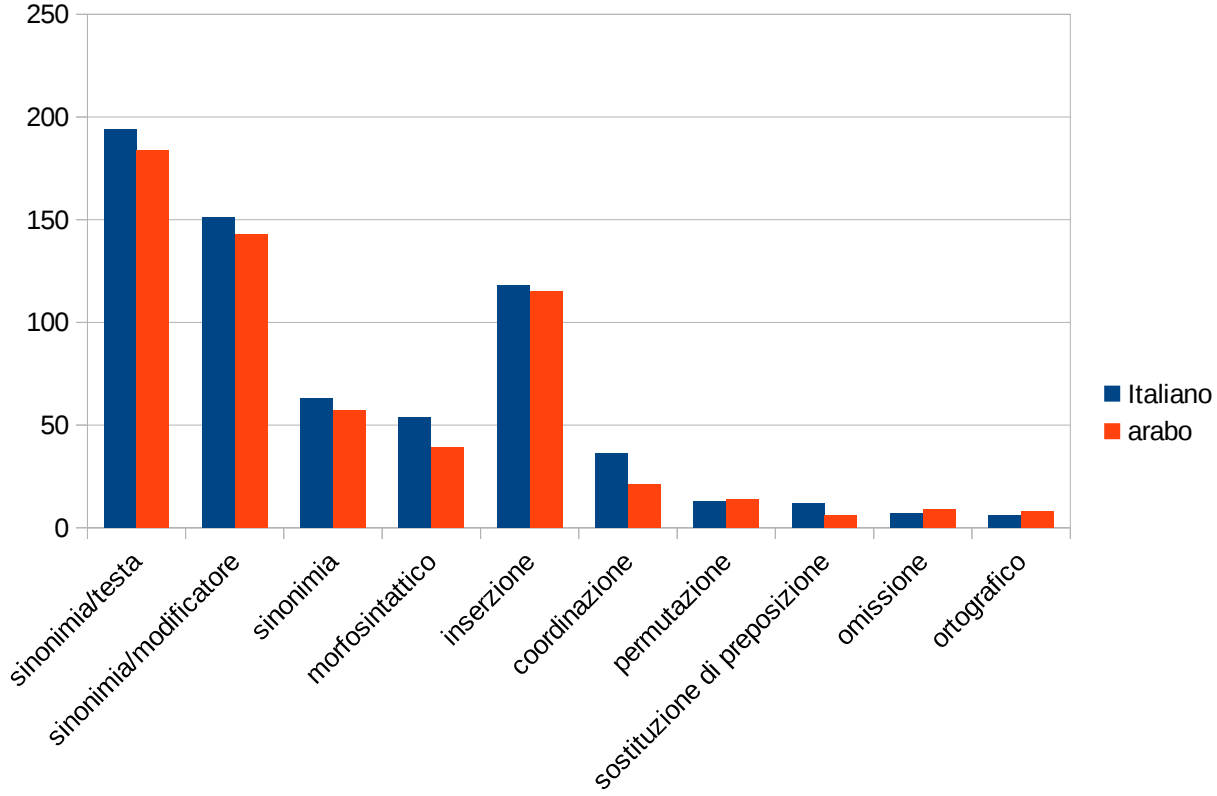
Termine di base arabo	Varianti del Termine di base arabo	n
Al>HkAm AlqAnwnyp disposizioni di legge	AlAHkAm AlqAnwnyp	1
Hqwq Al<nsAn diritti umani	Hqwq AlAnsAn	1
Al|vAr Alslbyp effetti negativi	AlAvAr Alslbyp	1
AlHq fY AxtyAr mdAfE diritto alla difesa	AlHq fy AxtyAr mdAfE	1
AlmdEy AlEAm procuratore generale	AlmdEY AlEAm	1
AlqAnwn AljnA}y diritto penale	AlqAnwn AljnA}Y	1
Altmyyz AlEnSry discriminazione razziale	Altmyyz AlEnSrY	1
>HkAm Al<EdAm pene di morte	AHkAm Al<EdAm	1
8		

6.4. Analisi dei dati

Dal numero totale di 1070 termini bilingui estratti nel capitolo precedente, ci sono 450 termini italiani che hanno presentato variazioni, mentre il numero dei termini arabi che hanno mostrato varianti è 458. La tabella seguente indica le diverse statistiche delle variazioni estratte per ogni lingua.

Variazione	Italiano			Arabo		
	N/termini	N/varianti	percentuale	N/termini	N/varianti	percentuale
sinonimia/ testa	194	298	43 %	184	270	40 %
sinonimia/ modificatore	151	225	33,5 %	143	195	31 %
sinonimia	63	84	14 %	57	77	12 %
morfo-sintattico	54	48	12 %	39	39	8,5 %
inserzione	118	145	26 %	114	136	24,9 %
coordinazione	36	41	8 %	21	22	4,6 %
permutazione	13	15	2,9 %	14	16	3 %
sostituzione di preposizione	12	14	2,6 %	6	6	1,3 %
omissione	7	7	1,5 %	9	9	1,9 %
ortografico	6	6	1,3 %	8	8	1,7 %

Tabella(26): Statistica delle variazioni nelle due lingue del corpus



Fig(18) diagramma della statistica delle variazioni nelle due lingue del corpus

Per le variazioni di sinonimia possiamo fare le seguenti osservazioni:

1- Stando al diagramma precedente si nota che la statistica più elevata riguarda la sinonimia con sostituzione di testa, seguita dalla sinonimia con variazione di modificatore, mentre il terzo posto viene occupato dalla sinonimia completa.

2- sia per la testa che per il modificatore, la relazione lessicale tra la componente originale e la variante ottenuta può essere uno di questi tipi:

- uno2uno_testa :
  - **servizio** militare ➡ **leva** militare
  - **mrswm** r}Asy (decreto presidenziale) ➡ **qrAr** r}Asy;

- uno2uno_ modificatore:
  - servizio **militare** ➡ servizio **nazionale**
  - <jAzp **AlwDE** (congedo per maternità)➡ <jAzp **Al>mwmp**;

- uno2due_ testa:
  - **matrimonio** omosessuale ➡ **unione civile** omosessuale
  - **AlqAnwn** AljnA}y (diritto penale) ➡ **nZAm AlEdAlp** AljnA}yp;

- uno2due_ modificatore:
  - indipendenza **della magistratura**➡ indipendenza **dell'autorità giudiziaria**
- Hryp **AltjmE** (libertà di associazione)➡ Hryp **tkwyn AljmEyAt**;

- due2uno_ testa:
  - **rapporti sessuali** illeciti ➡ **promiscuità** sessuale
  - **<sA'p AstxdAm** AlslTp (abuso d'ufficio)➡ **AstglAl** AlslTp;

- due2uno_ modificatore:
  - organizzazione **messa al bando** ➡ organizzazione **vietata**
  - mnZmAt **AlmjtmE Almdny**(organizzazioni della società civile) ➡ AlmnZmAt **Almdnyp**

3- In italiano in alcuni casi un sintagma preposizionale che funge da complemento o argomento del sintagma nominale può essere sostituito da due sintagmi preposizionali:
- libertà dei media ➡libertà dei mezzi di informazione;

4- In arabo il modificatore sostituto può essere composto da due elementi coordinati, come in
Alqtl AlEmd (omicidio colposo)➡ Alqtl mE sbq Al<SrAr wAltrSd

5- In caso di un modificatore poliematico, la sostituzione può riguardare o un solo costituente, come nei seguenti esempi:

-appartenenza a **gruppo** criminale ➡ appartenenza a **un'organizzazione** criminale
- AlAntmA' <lY **jmAEp** <jrAmyp ➡ AlAntmA' <lY **mnZmp** AjrAmyp;

- appartenenza a gruppo **criminale** ➡ appartenenza a gruppo **illegale**
- AlAntmA' <lY jmAEp **<jrAmyp** ➡ AlAntmA' <lY jmAEp **gyr qAnwnyp;**

oppure tutti gli elementi principali del modificatore, come in :
appartenenza a **gruppo criminale** ➡ appartenenza a **organizzazione fuorilegge**
- HSAnpF mn **AlmqADAp AljnA}yp** (immunità giudiziaria) ➡ AlHSAnp mn **AlmlAHqp AlqDA}yp**

6- la variante può subire cambiamento a livello dell'ordine tra la testa e il modificatore, come nel caso seguente:
- rappresentante legale ➡ avvocato rappresentante
- Al<dArp Alslymp (buona gestione) ➡ Hsn AdArp

7- la modifica di una componente può comportare modifica anche della preposizione che accompagna la nuova unità lessicale utilizzata:
- pena **di** morte ➡ condanna **a** morte
- AlHSwl **ElY** AltElym (accesso all'istruzione) ➡ AlHq **fy** AltElym;

oppure una cancellazione della preposizione presente nel termine originale dal momento che la nuova componente non richiede nessuna preposizione, come per es.
- >mr **b**AlqbD (mandato d'arresto) ➡ m*krp qbD;
o in casi contrari l'aggiunta di una preposizione:
mnZmAt AlmjtmE Almdny (organizzazioni della società civile) ➡ jmAEAt **mn** AlmjtmE Almdny

8- la variazione di una componente può comportare un cambiamento morfo-sintattico, come negli esempi:

- pena **di morte** ➡ pena **capitale**
- sistema **penitenziario** ➡ sistema **di prigione**

- qAnwn **AleqwbAt** (il diritto penale) ➡ qAnwn **AljnAy**
- Al>mn **AlwTny** (sicurezza nazionale) ➡ >mn **Aldwlp**;

Riassumiamo di seguito i più importanti pattern di variazione semantica incontrati nel corpus:

Pattern	Esempio
NpN ➡ **Np**N	**pena di** morte ➡ **condanna alla** morte
N**pN** ➡ N**A**	diritti **all'alloggio** ➡ diritti **abitativi**
N**pN** ➡ **N**A	prigioniero **di guerra** ➡ **soldato** prigioniero
NpN ➡ Np**NpN**	libertà di **stampa** ➡ libertà degli **organi d'informazione**
NpN ➡ Np**N**	libertà di **espressione** ➡ libertà di **parola**
NpN ➡ **AN**pN	**mancanza** di sicurezza ➡ **precarie condizioni** di sicurezza
NpN ➡ **ANpN**	**uguaglianza dinnanzi alla legge** ➡ **eguale trattamento avanti i tribunali**
NpN ➡ **NpN**	**disposizione di legge** ➡ **sensi della legislazione**
NpN ➡ **NApN**	**diniego della cittadinanza** ➡ **privazione arbitraria della nazionalità**
NpN ➡ **NA**	**emendamento alla legge** ➡ **cambiamenti legislativi**
NA ➡ **N**A	**grazia** presidenziale ➡ **aminstia** presidenziale
NA ➡ N**pA**	perquisizioni **corporali** ➡ perquisizioni **a nudo**
NA ➡ N**pNA**	lacune **legislative** ➡ lacune **del sistema giudiziario**
NA ➡ N**pN**	manifestazioni **antigovernative** ➡ manifestazioni **contro il governo**

NA ➡ N**A**p**N**	persone **disabili** ➡ persone **portatrici di handicap**
NA ➡ N**p**N**p**N	violenza **domestica** ➡ violenza **all'interno della famiglia**
NA ➡ N**A**	violenza **domestica** ➡ violenza **familiare**
N**A** ➡ N**p**N**p**A**N**	matrimonio **omosessuale** ➡ matrimonio **tra persone dello stesso sesso**
NA ➡ A**N**	**persone** eminenti ➡ eminenti **figure**
NA ➡ **NpAN**	**procuratore generale** ➡ **direttore della pubblica accusa**
NA ➡ **NA**	**funzionario pubblico** ➡ **agente statale**
NA ➡ **NpN**	**rimpatri forzati** ➡ **espulsioni di massa**
NA ➡ **NApN**	**salute professionale** ➡ **servizi sanitari sul lavoro**
NA ➡ **NpNA**	**popolazioni native** ➡ **gruppi di popoli indigeni**
NA ➡ **NAA**	**popolazioni native** ➡ **comunità indigena locale**
NA ➡ **NpNpN**	**detenzione pre-processuale** ➡ **carcere in attesa di giudizio**
AN ➡ **NpNpN**	**primo ministro** ➡ **presidente del consiglio dei ministri**
NAA ➡ **NAA**	**direzione** criminale investigativa ➡ **polizia** investigativa criminale
NA**A** ➡ NA**A**	rapporti sessuali **illeciti** ➡ rapporti sessuali **extraconiugali**
NAA ➡ **NApN**	**direzione criminale investigativa** ➡ **sezione investigativa della polizia**
NpNA ➡ **N**pNA	**organizzazioni** della società civile ➡ **associazioni** della società civile
NpN**A** ➡ NpN**A**	mezzi di **comunicazione** elettronici ➡ mezzi d'**informazione** elettronici
N**pNA** ➡ N**A**	rapina **a mano armata** ➡ rapina **aggravata**
NpN**A** ➡ NpN**A**	appartenenza a gruppo **criminale** ➡ appartenenza a un gruppo **illegale**
NpA**N** ➡ **N**pN**A**	**polizia** di rapido intervento ➡ **unità** di reazione rapida
NpA**N** ➡ NpA**N**	oltraggio a pubblica **autorità** ➡ oltraggio a pubblico **ufficiale**

NApN ➡ **N**ApN	**certificato** medico di attitudine ➡ **esame** medico di attitudine
NA**pN** ➡ NA**A**	consiglio supremo **della magistratura** ➡ consiglio supremo **giudiziario**
N**A**pN➡N**A**pN	consiglio **supremo** della magistratura ➡ consiglio **superiore** della magistratura
NpNpN ➡ **N**pN	**sentenze con sospensione** della pena ➡ **abolizione** della pena
NpAN ➡ **NpNA**	**polizia di rapido intervento** ➡ **forza d'intervento speciale**
NpAN ➡ **NpAN**	**oltraggio a pubblico ufficiale** ➡ **vilipendio a pubblica autorità**

Tabella (27). Pattern delle varianti semantiche del corpus italiano

Pattern	Esempio
NN ➡ **N**N	**Hkm** Al<EdAm (pena di morte) ➡ **Eqwbp** Al<EdAm
NN ➡ **NpN**N	**tzwyr** AlAntxAbAt (frodi elettorali) ➡ **AltlAEb bntA}j** AlAntxAbAt
NN ➡ NN**A**	wzyr **AlEdl** (ministro della giustizia) ➡ wzyr **Al$&wn AlqAnwnyp**
NN ➡ NN**N**	Hryp **AlSHAfp** (libertà di stampa) ➡ Hryp **wsA}l Al<ElAm**
NN ➡ N**ApNA**	EA}dAt **Aljrymp** (proventi del reato) ➡ AlEA}dAt **Almt>typ mn Al>fEAl Almjr~mp**
NN ➡ **NAp**N	**AxtSAS** AlmHkmp (competenza della corte) ➡ **AlwlAyp AlqDA}yp l**mHAkm
NN ➡ **NN**	**>HkAm Alsjn** (sentenza di carcere) ➡ **Eqwbp AlHbs**
NN ➡ **NA**	**tEdyl qAnwn** (modifica di legge) ➡ **AltgyyrAt Alt$ryEyp**
NN ➡ **NNN**	**fsx Eqd** (annullamento di un contratto) ➡ **<nhA' AtfAq AlAstxdAm**
NN ➡ **NpN**	**tzwyr AlAntxAbAt** (brogli elettorali) ➡ **tlAEb fy Al>SwAt**
NA ➡ **N**A	**mrswm** r}Asy (decreto presidenziale) ➡ **qrAr** r}Asy

NA ➡ NA	AlslTp **AlmxtSp** (autorità competente) ➡ AlslTp **Alms}wlp**
NA ➡ NN	mwZf **Emwmy** (ufficiale pubblico) ➡ mwZfy **AlHkwmp**
NA ➡ NNA	**jnsyp** mzdwjp (doppia cittadinanza) ➡ **jwAz sfr** mzdwj
NA ➡ NNA	AlmnZmAt **AltTwEyp** (organizzazioni umanitarie) ➡ mnZmAt **Al<gAvp Al<nsAnyp**
NA ➡ N**pN**	mwZf **Emwmy** (ufficiale pubblico) ➡ mwZfyn **fy Aldwlp**
NA ➡ N**pNN**	AlEnf **Al>sry** (violenza domestica) ➡ AlEnf **fy <TAr Al>srp**
NA ➡ N**AA**	AlEmAl **AlmhAjryn** (lavoratori immigrati) ➡ AlEmAl **Al>jAnb AlwAfdyn**
NA ➡ N**ApNNA**	AlmnZmAt **AltTwEyp** (organizzazioni umanitarie) ➡ AlmnZmAt **AlEAmlp fy mjAl AlmsAEdAt Al<nsAnyp**
NA ➡ N**pNA**	zwAj **mvly** (matrimonio omosessuale) ➡ zwAj **mn Aljns AlwAHd**
NA ➡ N**pNpAN**	zwAj **mvly** (matrimonio omosessuale) ➡ AlzwAj **byn >frAd mn nfs Aljns**
NA ➡ N**AN**	Alr}ys **AlmnSrf** (presidente uscente) ➡ Alr}ys **Almnthyp wlAyth**
NA ➡ NA	**>qlyAt Erqyp** (minoranza etnica) ➡ **AlmjtmEAt Al<vnyp**
NA ➡ NN	**AlAbEAd Alqsry** (rimpatrio forzato) ➡ **EmlyAt AltrHyl**
NA ➡ NNN	**AlmdAfE AlEAm** (difensore civico) ➡**>myn dywAn AlmZAlm**
NA ➡ NAA	**Alxdmp AlEskryp** (leva militare) ➡ **Altjnyd AlwTny Al<lzAmy**
NA ➡ NpNA	**Alt|mr AljnA}y** (cospirazione criminale) ➡ **AlArtbAT mE EnASr <jrAmyp**
NA ➡ NpNN	**Alt|mr AljnA}y** (cospirazione penale) ➡ **AltwAT& lArtkAb jrymp**
NNN ➡ NNN	**qAnwn** mkAfHp Al<rhAb (legge di antiterrorismo) ➡ **t$ryE** mkAfHp Al<rhAb

N**NN** ➡ N**NN**	qAnwn **mkAfHp Al<rhAb** (legge di antiterrorismo) ➡ qAnwn **mnE Al<rhAb**
NNN ➡ **NNA**	**$rTp mkAfHp Al$gb** (polizia antisommossa) ➡ **qwAt Al>mn Almrkzy**
NNN ➡ **NN**	**<sA'p AstxdAm AlslTp** (abuso dell'ufficio) ➡ **AstglAl AlmnSb**
NNA ➡ **N**NA	**$rTp** Altdxl AlsryE (polizia di rapido intervento) ➡ **qwp** Altdxl AlsryE
N**N**A ➡ N**N**A	n$r **>xbAr** kA*bp (pubblicazione di notizie false) ➡ n$r **mElwmAt** kA*bp
N**NA** ➡ N**NA**	<dArp **AltHqyqAt AljnA}yp** (direzione criminale investigativa) ➡ <dArp **AlbHv AljnA}y**
N**N**A ➡ N**A**A	qAnwn **AljrA}m** Aldwlyp (legge penale internazionale) ➡ AlqAnwn **AljnA}y** Aldwly
N**N**A ➡ N**NN**A	qAnwn AljrA}m Aldwlyp (legge penale internazionale) ➡ qAnwn **mHkmp AljnAyAt** Aldwlyp
N**N**A ➡ N**N**A	qAnwn **AljrA}m** Aldwlyp (legge penale internazionale) ➡ qAnwn **AlmHAkm** Aldwlyp
NNA ➡ **NNp**NA	**AtlAf** AlmmtlkAt AlEAmp (danneggiamento delle proprietà pubbliche) ➡ **AlHAq AlDrr b**AlmmtlkAt AlEAmp
NNA ➡ **Np**NA	**mnZmAt** AlmjtmE Almdny (organizzazioni della società civile) ➡ **jmAEAt mn** AlmjtmE Almdny
NNA ➡ **NNA**	**$rTp Altdxl AlsryE** (polizia di rapido intervento) ➡ **fylq AltHrk AlsryE**
NpN ➡ **N**pN	**Altjryd** mn Aljnsyp (diniego della cittadinanza) ➡ **AlHrmAn** mn Aljnysp
NpN ➡ **N**N	**Amr b**AlqbD (ordine di arresto) ➡ **m*krp** qbD
N**pN** ➡ N**pNA**	AlmsAwAp **byn Aljnsyn** (uguaglianza di genere) ➡ AlmsAwAp **fy AlnwE AlAjtmAEy**
N**pN** ➡ N**ApNN**	AlmsAwAp **byn Aljnsyn** (uguaglianza di genere) ➡ AlmsAwAp **AlqA}m ElY nwE Aljns**

Np**N** ➡Np**NNN**	AlHq fy **AlAst}nAf** (diritto di appello) AlHq fy **rfE** ➡ **dEAwY Ast$kAl**
Np**N** ➡ Np**NpN**	AlHq fy **AldfAE** (diritto alla difesa) ➡ AlHq fy **AlAstEAnp bmHAmy**
N**pN** ➡ N**NA**	AlHq **fy AltnZym** (diritto sindacale) ➡Hq **AltnZym AlnqAby**
N**pN** ➡ N**A**	AlHq **fy AltnZym** (diritto sindacale) ➡ AlHq **AlnqAby**
NpN ➡ **NpN**N	**AlHSwl Ely** AltElym (accesso all'istruzione)➡ **Al<ltHAq bm&ssAt** AltElym
NpN ➡ **NN**	**>mr bAlqbD** (ordine di arresto) ➡ **m*krp AetqAl**
NpN ➡ **NA**	**AlmsAwAp fy AlmEAmlp** (trattamento eguale) ➡ **AlmEAmlp AlEAdlp**
NAN ➡ **N**AN	**Almn$A'At** mtEddp AljnsAt (imprese multinazionali) ➡ **Al$rkAt** mtEddp AljnsAt
NpNA➡ **N**NA	**qSwr fy** AlnZAm AlqDA}y (disfunzione nel sistema giudiziario) ➡ **tqAEs** AlnZAm AlqDA}y
Np**N**A➡ Np**N**A	AltEdy ElY **ms&wl** EAm (oltraggio a pubblico ufficiale) ➡ AltEdy ElY **mwZf** Eam
Np**NA**➡ Np**NA**	HSAnpF mn **AlmqADAp AljnA}yp** (immunità giudiziaria) ➡ HSAnp mn **AlmlAHqp AlqDA}yp**
NpNA➡**NpN**NA	**AltEdy Ely** ms}wl EAm (oltraggio a pubblico ufficiale) ➡ **Alqzf fy Hq** ms}wl EAm
NpNA➡ **NNA**	**AltEdy ElY ms&wl EAm** (oltraggio a un pubblico ufficiale)➡ **<hAnp mwZfyn rsmyyn**
Np**N**N➡ N**pN**N	AltSwyt **ElY Hjb** Alvqp (voto di sfiducia)➡AltSwyt **bsHb** Alvqp
NApN➡ **NA**pN	**AlmZAhrAt AlmEAdyp** llHkwmp (manifestazioni antigovernative) ➡**AlmsyrAt AlmnAhDp** llHkwmp
N**A**pN➡ N**A**pN	AlqwAt **AlmwAlyp** llHkwmp (forze filogovernative) ➡ AlqwAt **Alm&ydp** llHkwmp
N**ApN**➡ N**ApN**	AlqwAt **AlmwAlyp** l**lHkwmp** (forze filogovernative) ➡ AlqwAt **AltAbEp** l**lnZAm**

N**ApN** ➡ N**A**	AlAHtjAz **AlsAbq llmHAkmp** (detenzione pre-processuale) ➡ AlAHtjAz **AlAHtyATy**
NApN ➡ **NA**	**AlAHtjAz AlsAbq llmHAkmp** (detenzione preprocessuale) ➡ **AlHbs AlAHtyATy**

Tabella (28). Pattern delle varianti semantiche del corpus arabo

Per le variazioni morfo-sintattiche e sintattiche possiamo fare le seguenti osservazioni:

1-Nelle variazioni morfo-sintattiche si trovano i seguenti pattern:

N**A** ➡ N**p**N:
campagna **elettorale** ➡ campagna **per le elezioni**

N**A** ➡ N**A**
rilascio **condizionato** ➡ rilascio **condizionale**

N**p**N ➡ N**N**:
-assicurazione **per la malattia** ➡ assicurazione **malattia**

N**p**N ➡ N**A**:
- disposizione **di legge** ➡ disposizione **legale**

N**A** ➡ N**p**N:
- tEdylAt **dstwryp (**riforme costituzionali) ➡ tEdylAt **ElY Aldstwr**

N**A** ➡ N**N**:
-AlqAnwn **AljnA}y** (codice penale) ➡ qAnwn **AljnAyAt**
N**N** ➡ N**A**:
tlwv **Alby}p** (inquinamento ambientale) ➡ Altlwv **Alby}y**

Una variazione importante in questa categoria è quella derivazionale, come nei casi seguenti:

NA ➡ **N**A :
- **crimine** organizzato ➡ **criminalità** organizzata

N**A** ➡ N**A**:
- stereotipi **sessisti** ➡ stereotipi **sessuali**

N**A** ➡ N**pN**
persone **disabili** ➡ persone **con disabilità**

N**pN** ➡ N**A**
azioni **di violenza** ➡ azioni **violente**

NN ➡ **N**N
Eml Al>TfAl (lavoro minorile) ➡ **EmAlp** Al>TfAl

N**N** ➡ N**N**
Hqwq **Alskn** (diritto all'alloggio) ➡ Hqwq **AlmskAn**

N**A** ➡ N**A**
AlA}tlAf **AlHAkm**(coalizione di governo) ➡ AlA}tlAf **AlHkwmy**

NA ➡ **N**A :
-**AlnzAEAt** AlmslHp (conflitti armati) ➡ **AlmnAzEAt** AlmslHp

NA ➡ N**A**A:
- **Aljrymp** AlmnZmp (crimine organizzato) ➡ Aln$AT **Al<jrAmy** AlmnZm (attività criminali organizzate)

In alcuni casi si possono trovare compresenti più variazioni all'interno dello stesso termine, come per es.

- variazione morfo-sintattica con sostituzione di una testa
grazia presidenziale ➡ clemenza del presidente

Almrswm Alr}Asy (decreto presidenziale) ➡ qrAr Alr}ys

- variazione morfo-sintattica con mutazione
inchiesta indipendente ➡ indipendenza dell'inchiesta
AstqlAl AlqDA' (indipendenza della magistratura) ➡ AlqDA' Almstql

- variazione morfo-sintattica con inserzione
età lavorativa ➡ età minima per il lavoro
AlHq fy AldfAE (diritto alla difesa) ➡ AlHq fY AxtyAr mdAfE (diritto alla scelta di un difensore)

- variazione morfo-sintattica con omissione
reato di matrice razziale ➡ reato razzista
AlwlAyp AlqDA}yp AlEskryp (giurisdizione militare) ➡AlqDA' AlEskry

2- La variazione del tipo sintattico può essere accompagnata anche dalla modifica di uno o più dei costituenti principali del termine, per cui abbiamo classificato ogni tipo di variazione sintattica in due categorie: variazioni senza modifica delle componenti originali e variazioni con modifica delle componenti del termine.

Le variazioni di inserzione riguardano le modifiche dovute all'inserzione di altri elementi lessicali all'interno delle componenti del termine originale. Questi elementi possono essere aggettivi, nomi o preposizioni.

- inserzione di un aggettivo
autorità competente ➡ autorità **amministrativa** competente
AlslTp AlmxtSp (autorità competente) ➡ AlslTp **AlqDA}yp** AlmxtSp (AlqDA}yp = giudiziale)

- inserzione di un nome
personale carcerario ➡ personale **del servizio** carcerario
m$rwE qAnwn (progetto di legge)➡m$rwE **tEdyl** qAnwn(tEdyl=modifica)

- inserzione di un sintagma preposizionale
misure legislative ➡ misure **in campo** legislativo
AlEnf fy mHyT Al>srp (violenza familiare) ➡ AlEnf **Dd Almr>p** fy mHyT Al>srp (**Dd Almr>p = contro donna**)

- inserzione di una preposizione
salario base ➡ salario **di** base
Hkm Al<EdAm (condanna a morte) ➡ Hkm **b**Al<EdAm (b = con)

- inserzione di due elementi lessicali
camera d'appello ➡ camera **penale della corte** d'appello
mnZmp Hqwq Al<nsAn (organizzazione dei diritti umani) ➡ mnZmAt **mHlyp mEnyp** bHqwq Al<nsAn (mHlyp mEnyp = locali e interessati)

Per quanto concerne le variazioni di inserzione con modifica delle componenti del termine, qui la sostituzione può riguardare sia la testa che il modificatore del termine, come dimostrano gli esempi seguenti:

- diritti **umani** ➡ diritti **fondamentali della persona**
-<jAzp **AlwDE** (congedo di maternità) ➡ <jAzp **<lzAmyp bEd AlwlAdp** (congedo di maternità obbligatoria)
- violenza sessuale ➡ **abuso di natura** sessuale
mnZmAt AlmjtmE Almdny (organizzazioni della società civile) ➡ **jmAEAt mn** AlmjtmE Almdny (gruppi della società civile)

3- Per quanto riguarda le variazioni sintattiche dovute a motivi di coordinazione possiamo notare che anche in questo caso la coordinazione può essere accompagnata da sostituzione di un costituente del termine originale. In entrambi i casi, cioè con o senza sostituzione di un elemento principale del termine, la coordinazione può avvenire tra la testa o il modificatore del termine e un nuovo elemento inserito, come dimostrano gli esempi seguenti:

- uso eccessivo della forza ➡ uso eccessivo e **indebito** della forza

- >qlyAt Erqyp (minoranza etnica) ➡ Al>qlyAt **AlEnSryp** wAlErqyp (AlEnSryp = razziale)
- matrimonio omosessuale ➡ matrimonio o **unione** omosessuale
- Alt>hyl Almhny (formazione professionale) ➡ **Altdryb** wAlt>hyl Almhny (Altdryb = addestramento)

4-anche le variazioni per permutazione possono subire modifiche degli elementi principali del termine. In questo caso la permutazione consiste nello scambiare l'ordine di successione tra la testa il modificatore del termine:

- possesso illegale di armi ➡ possesso di armi illegale
- Almjls Al>ElY llqDA' (Consiglio supremo della magistratura) ➡ mjls AlqDA' Al>ElY
- uso eccessivo della forza ➡ impiego di forza eccessiva
->srY AlHrb (prigionieri di guerra) ➡ Aljnwd Al>srY

5- Per quanto riguarda le variazioni sintattiche possiamo notare che entrambe le lingue presentano un'elevata percentuale delle variazioni per motivi di inserzione, seguita dalla coordinazione e poi viene la permutazione. Per gli ultimi due posti si nota una differenza tra le due lingue: mentre in italiano la sostituzione di preposizione occupa il quarto posto e l'omissione viene all'ultimo posto, in arabo le variazioni dovute all'omissione precedono quelle per sostituzione di preposizione.

6- Per quanto riguarda le variazioni ortografiche possiamo notare in generale che in italiano i casi registrati dimostrano che la maggior parte delle varianti ortografiche sono dovute alla presenza o l'assenza dell'apostrofo dinanzi alle parole che iniziano con vocale, oppure all'uso o meno del trattino con i nomi composti, come in *anti-terrorismo* o *antiterrorismo*. Le variazioni ortografiche osservate nel corpus arabo si possono classificare principalmente in tre categorie: a) la presenza o meno di *hamza* (ء) che si mette sopra o sotto *alif* (A), come in *Al>HkAm* (disposizioni) o *AlAHkAm*; b) la presenza o la mancanza di *madda* (~) sopra *alif* (A), come in, *Al|vAr* (effetti) o *AlAvAr*; c) l'uso di *ya* finale con o senza i due puntini, come *fy* (in) o *fY*.

Conclusione

In questo lavoro ho cercato di utilizzare gli strumenti del TAL per estrarre le variazioni semantiche, sintattiche e morfosintattiche delle terminologie poliriematiche in un corpus giuridico parallelo italiano- arabo. La ricerca si serve dell'approccio computazionale nelle diverse fasi del lavoro: la creazione e l'annotazione del corpus parallelo, l'estrazione dei termini dai corpora monolingui, l'allineamento dei termini bilingui, e l'identificazione delle variazioni terminologiche. Nell'ambito del percorso della ricerca non mancava, tuttavia, qualche intervento manuale, finalizzato principalmente a verificare i dati selezionandone solo i candidati significativi e rilevanti. Le maggiori sfide affrontate durante il lavoro riguardano effettivamente l'applicazione dell'approccio computazionale, opzione non priva di difficoltà considerando le differenze linguistiche delle lingue del corpus nonché lo stato dell'arte del TAL arabo che, malgrado il recente interesse da parte dei ricercatori linguistico- computazionali, non dispone ancora di strumenti sofisticati per analizzare i testi soprattutto a livello sintattico. Un'enorme sfida fronteggiata dalla ricerca era la possibilità di avere a disposizione testi giuridici paralleli italiano-arabo.

Nel primo capitolo della tesi ho cercato di evidenziare teoricamente il fenomeno delle variazioni terminologiche, partendo dalla concezione tradizionale della terminologia, basata sul principio di monoreferenzialità dei termini per cui i termini non devono presentare variazioni a livello del testo, passando poi per gli approcci socio-cognitivi e testuali che pongono al centro della loro riflessione la dimensione socio-cognitiva e comunicativa della terminologia, superando i limiti della visione prescrittiva della teoria tradizionale *wusteriana.* Pur avendo differenze nel modo di studiare la terminologia, le nuove scuole di pensiero hanno in comune tante caratteristiche tra cui, oltre al rifiuto del metodo prescrittivo classico, l'adozione dell'approccio descrittivo che prova a estrarre evidenze empiriche fornite da corpora testuali, nonché l'utilizzo dell'approccio semasiologico che può essere integrato con alcuni contributi onomasiologici. Dopo il rifiuto del distacco tra la semantica dei linguaggi specializzati e la loro veste linguistica,

questi approcci descrittivi della terminologia dimostrano come la variazione denominativa non riguardi solo la lingua generale, ma anche i discorsi specializzati, contrastando perciò con il principio di univocità dei messaggi tecnico-scientifici. L'ammissione della variazione terminologica nelle lingue speciali comporta anche la revisione del concetto dell'equivalenza tra i termini, che secondo alcuni studiosi dovrebbe essere vista come equivalenza parziale, dal momento che la nozione di equivalenza assoluta, soprattutto fra i termini giuridici che hanno una certa peculiarità dovuta all'ancoraggio e al radicamento culturali, appare molto difficile.

Prima di procedere all'estrazione di termini dal corpus era importante soffermarci nel secondo capitolo sui procedimenti della formazione delle parole in italiano e in arabo, fase indispensabile per riconoscere le caratteristiche linguistiche dei termini giuridici in entrambe le lingue, che, in genere, sono soggetti agli stessi procedimenti formativi dei lessici nella lingua comune. Questa parte del lavoro ha messo in risalto alcune differenze tra le due lingue riguardo alla formazione dei loro lessici: mentre la lingua italiana dipende maggiormente dal procedimento di composizione per rinnovare e arricchire i suoi vocaboli, nella lingua araba la derivazione si configura come la fonte principale della formazione delle parole. Questo non impedisce, però, la presenza di punti di affinità in relazione alla formazione dei termini tecnici: un processo condiviso da ambedue le lingue per formare nuovi tecnicismi è la terminologizzazione. Si tratta o di specializzazione di un'unità lessicale, quando cioè una parola proviene dalla lingua comune e passa a un linguaggio specializzato acquisendo un'entità semantica differente da quella che possiede nel suo contesto generale, o tramite il prestito di termini da lingue straniere.

Partendo dall'assunto che analizzare le terminologie tramite il contesto linguistico e situazionale aiuti a capire il funzionamento reale dei termini, abbiamo deciso di studiare le variazioni terminologiche per mezzo di un corpus linguistico. Per questo scopo è stato costituito nel terzo capitolo un corpus parallelo italiano-arabo nel dominio dei diritti umani nel mondo. Si sono raccolti, quindi, accordi, convenzioni, protocolli, rapporti internazionali riguardanti la situazione dei diritti umani nel mondo.

Una volta riconosciute le caratteristiche morfo-sintattiche dei termini nelle due lingue, procediamo, nel quarto capitolo, a formare dei pattern linguistici tipici delle unità terminologiche in italiano e in arabo per estrarre inizialmente dei termini candidati che vengono filtrati successivamente tramite i metodi statistici. Le misure statistiche adottate per ottimizzare i dati si dividono in due classi: metodi di *unithood*, che misurano quantitativamente il grado di associazione di combinazioni sintagmatiche o di collocazioni; e metodi di *termhood* che si riferiscono al grado di rilevanza o significatività di un'unità linguistica nei riguardi di un concetto nell'ambito di un dominio specifico. Nel caso dei metodi di *unithood* si sono utilizzati i due test Log Likelihood Ratio e Mutua Informazione, mentre per la *termhood* ci siamo serviti della funzione di CN-value. Confrontando i risultati dell'estrazione dei termini monolingue si può notare qualche differenza, a favore dei testi italiani, nella performance delle misure statistiche, tasso di differenza che si può attribuire all'accuratezza nel compito di PoS tagging nonché alla complessità del sistema linguistico arabo.

I termini rilevanti estratti nella fase precedente rappresentano poi il nucleo dell'estrazione bilingue finalizzata alla creazione di corrispondenze di traduzione, che è il tema del quinto capitolo. Essenziale in questo stadio del lavoro è il ruolo del corpus parallelo che si utilizza per consentire la ridistribuzione contestuale dei termini individuati che verranno poi riestratti dal contesto parallelo a livello di unità di traduzione. Dopo la verifica delle relazioni positive, le unità cioè che non contengano elementi vuoti sia nella SL che nella TL, si procede a convalidare delle equivalenze traduttive, utilizzando un approccio statistico-linguistico. La parte statistica dell'approccio consiste nell'utilizzo del test LLR e di un sistema di traduzione automatica statistica (SMT). L'idea base del LLR parte dal fatto che le statistiche fornite dal corpus parallelo per ogni paio di termini bilingui possono essere utilizzate per creare una tabella di contingenza tramite cui si misura il valore di traduzione o di co-occorrenza tra i termini bilingui. Usare un SMT viene suggerito, invece, dall'ipotesi che tramite il confronto tra un termine composto e la sua relativa traduzione in un'altra lingua si possano identificare termini bilingui equivalenti. La parte linguistica dell'approccio

comprende la funzione della posizione dei termini all'interno delle frasi. Questo metodo si basa sulla teoria secondo la quale in due lingue come l'italiano e l'arabo, anche se possiedono un sistema libero dell'ordine delle parole a livello di frase, le parole, specialmente le unità polirematiche, si presentano probabilisticamente vicine a livello di posizione nella frase. Quest'ultima opzione di usare la posizione dei termini nel contesto parallelo appare conveniente nel nostro caso considerato il numero limitato dei termini bilingui da allineare a livello di ogni unità di traduzione. La valutazione dell'estrazione bilingue dimostra un f-measure alto nel caso del metodo statistico rispetto a quello linguistico.

Nell'ultimo capitolo del lavoro si procede a individuare le diverse forme di variazione delle unità terminologiche bilingui. In questa fase ci si serve delle seguenti informazioni: i termini bilingui, i termini candidati, i lemmi sia dei termini selezionati che dei termini candidati, il corpus parallelo, WordNet. In una prima fase si adotta la conformità dei lemmi tra i termini candidati e quelli selezionati insieme al concetto della sostituibilità dei termini all'interno del corpus parallelo. In una fase successiva si utilizzano i synset, forniti da WordNet in lingua inglese, per individuare relazioni di variazione, soprattutto semantiche, tra i termini selezionati e quelli candidati. I risultati delle variazioni semantiche dimostrano che la statistica più elevata riguarda la sinonimia con sostituzione di testa, seguita dalla sinonimia con variazione di modificatore, mentre al terzo posto viene collocata la sinonimia completa. Per le variazioni sintattiche e morfo-sintattiche notiamo che sia in italiano che in arabo la variazione dovuta all'inserzione è relativamente elevata e al secondo posto viene messa la variazione morfosintattica. Seguono, con percentuali vicine, la coordinazione, la permutazione, la sostituzione e l'omissione.

Pensiamo che il contributo principale del nostro lavoro, malgrado alcune limitazioni riguardanti soprattutto l'esclusione dei termini monorematici e di alcuni tipi di variazione come le parafrasi, può riguardare i seguenti punti:
- constatare tramite un approccio descrittivo la presenza del fenomeno della variazione denominativa nei linguaggi specialistici in generale e nel dominio giuridico in particolare;

- creare un corpus parallelo italiano-arabo nel campo del diritto internazionale, fornendo risorse linguistiche utili alle applicazioni della linguistica computazionale, soprattutto tra due lingue diverse come l'italiano e l'arabo;

- utilizzare un metodo ibrido che combina le informazioni linguistiche con i calcoli statistici per estrarre unità terminologiche monolingue da corpora di dominio, che possono servire in tante applicazioni nel campo linguistico;

- presentare nuove tecniche per estrarre delle corrispondenze di traduzione da corpora paralleli, che risultano di fondamentale importanza per arricchire i sistemi di traduzione automatica con nuove risorse lessicali. Tali tecniche possono contribuire a migliorare la performance dei lavori di allineamento a livello di parola sopratutto tra lingue con sistemi linguistici diversi;

- adottare un metodo innovativo per estrarre le variazioni terminologiche da corpora di dominio. Si tratta di un metodo che non solo impiega un contesto interlinguistico per verificare la sostituibilità dei termini, bensì si serve di un sistema semantico-lessicale come Wordnet, in una lingua pivot, per individuare le diverse variazioni dei termini monolingui.

- Bibliografia

- Aguado de Cea, G., Montiel-Ponsoda, E., "Term variants in ontologies". In *Proceedings of the 30th International Conference of AESLA*, Spain, 2012

- Al Khatib, K., Badarneh, A., "Automatic extraction of Arabic multi-word terms". In *Proceedings of the International Multiconference on Computer Science and Information Technology*, 2010

-Al-Taani A.T, Abu-Al-Rub S., "A rule-based approach for tagging non-vocalized Arabic words". In *The International Arab Journal of Information Technology,* Volume6 (3), 2009

- Alhaj, A., *Understanding Semantics. A Textbook for Students of Linguistics and Translation*, Hamburg, Anchor Academic Publishing, 2015

-Antia, B.E., *Terminology and Language Planning: An alternative framework of practice and discourse,* Amsterdam: John Benjamins Publishing Company, 2000

- Arntz R., "Terminological Equivalence and Translation". In *Terminology. Applications in Interdisciplinary Communication*, Sonneveld H. B.&Loening K. L. (eds), Amsterdam, John Benjamins, 1993

- Attia, M. et al., "Automatic Extraction of Arabic Multiword Expressions". In *Proceedings of the Workshop on Multiword Expressions: from Theory to Applications*, Beijing, China. 2010

- Baker, M., "Corpora in Translation Studies. An Overview and Suggestions for Future Research". In *Target*, 7(2). 1995

- Bannard, C., Callison-Burch, C., "Paraphrasing with bilingual parallel corpora". In *Proceedings of the annual Meeting of the Association for Computational Linguistics (ACL)*, Ann Arbor, MI, 2005

- Bar, K., Dershowitz, N., “Using semantic equivalents for Arabic-to-English Example-based translation”. In *Challenges for Arabic Machine Translation*, John Benjamins Publishing Company, 2012

- Barbagianni C., “Verso un modello di variazione terminologica: un’analisi della terminologia della gestione dei rifiuti in testi normativi”. In *Quaderni di Palazzo Serra* 25 (2014), 7-24.

- Baroni, E. et al., “The dual nature od deverbal nominal constructions: Evidence from acceptability ratings and corpus analysis”. In *Corpus Linguistics and Linguistic Theory* 5(1), 2009

- Basili, R., et al., "Customizable Modular Lexicalized Parsing", in *Proceedings of the 6th International Workshop on Parsing Technology*, IWPT, 2000

- Basili, R., et al., “A contrastive approach to term extraction”. In *Proceedings of the 4th Terminology and Artificial Intelligence Conference (TIA)*, France, 2001

– Basili, R., et al., “A text classifier based on linguistic processing”. In *Proceedings of IJCAI 1999. Machine Learning for Information Filtering*, 1999

- Belvedere, A., “Il linguaggio del codice civile: alcune osservazioni”. In Scarpelli, U., Di Lucia, P. (a cura di) *Il linguaggio del diritto*, Milano, LED, 1994

- Biber, D., “Representativeness in Corpus Design”. In *Literary and Linguistic Computing*, Volume 8, n.4, 1993

- Bisetto, A., “Note sui composti VN dell’italiano”. In Benincà, P. et al. (a cura di), *Fonologia e morfologia dell’italiano e dei dialetti d’Italia*, Atti del XXXI Congresso internazionale di studi della Società di Linguistica Italiana, Roma, Bulzoni, 1999

- Bolasco, S., “Statistica testuale e text mining: alcuni paradigmi applicativi”. In *Quaderni di statistica*, vol.7, 2005

- Bolasco, S., et al., “Estrazione automatica d'informazione dai testi”, in *Mondo digitale*, n.1 marzo 2004

- Bonin, F., et al. “A contrastive approach to multi word term extraction from domain corpora” In *Proceedings of the 7th International Conference on Language Resources and Evaluation* (LREC 2010), La Valletta, Malta

- Boulanger, J.C., “Présentation: images et parcours de la socioterminologie”. In *Méta*, XL, 2, 1995

- Bounhas I., Slimani,Y., “A hybrid approach for Arabic multi-word term extraction”. In *International Conference on Language Processing and knowledge Engineering*, 2009

- Bowker, L., “Variant terminology: frivolity or necessity?” In 8th *EURALEX International Congress*. Liège: University of Liège, 1998

- Bowker, L., Pearson, J., *Working with Specialized Languages: A practical guide to using corpora*, London and New York, Routledge, 2002

- Brew C., McKelvie, D.,“Word-pair extraction for lexicography”. In *Proceedings of International Conference on New Methods in Natural Language Processing*, Bilkent, Turkey, 1996

- Brown, P., et al., “A statistical approach to machine translation”. In *Computational Linguistics*, 16(2):1990, MA, USA, MIT Press Cambridge

- Cabré, M. T., “El principio de poliedricidad: la articulación de lo discursivo, lo cognitivo y lo lingüístico en Terminología”. In *IBÉRICA* 16, 2008
- Cabré, M. T., “Theories of terminology, their description, prescription and explanation’. In *Terminology* 9(2), 2003

- Cabré, M. T., *La terminología: Representación y comunicación. Elementos para una teoría de base comunicativa y otros artículos*, Barcelona, Institut Universitari de Lingüística Aplicada, 1999

- Cabré, M. T., *Terminology: theory, methods and applications*, Amsterdam: John Benjamins Publishing Company, 1998

- Cabré, M.T., "Standardization and Interference in Terminology". In *The Changing Scene in World Languages. Issues and challenges*, Amsterdam/Philadelphia, John Benjamins Publishing Company, 1997

- Callison-Burch, C.,"Syntactic Constraints on Paraphrases Extracted from Parallel Corpora". In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii. Association for Computational Linguistics.

- Carpuat, M. et al., "Improved Arabic-to-English statistical machine translation by reordering post-verbal subjects for word alignment". In *Proceeding ACLShort '10 Proceedings of the ACL* 2010

- Carreño Cruz, S., I., *Analyse de la variation terminologique en corpus parallèle anglais-espagnol et de son incidence sur l'extraction de termes bilingue*, tesi di master, Université de Montréal, 2004.

- Cavagnoli, S., *La comunicazione specialistica*, Roma, Carocci, 2007

- Chakraborty, R., "Domain Keyword Extraction Technique: a new weighting method nased on frequency analysis". In Bhattacharyya, R. et al. (a cura di):*ACER*, 2013

- Chiari, I., "La chiave probabilistica delle lingue: teoria linguistica e applicazioni computazionali". In Fulci, L. e Sciubba, E. (a cura di), *Linguaggio, Mente e Società*, Roma, EuRoma-La Goliardica, 2008

- Chiari, I., *Introduzione alla linguistica computazionale*, Roma, Editore Laterza, 2007

- Choueka, Y., "Looking for needles in a haystack or locating interesting collocational expressions in large textual databases". In *Proceedings of the RIAO Conference on User-Oriented Content-Based Text and Image Handling*, Cambridg, MA, 1988

- Condamines, A., "Variations in terminology: Application to the management of risks related to language use in the workplace". In *Terminology*; 2010, Vol. 16 Issue 1

- Cortelazzo, M.A., *Lingue speciali. La dimensione verticale*, Padova, Unipress, 1994

- Daille B., et al. "Towards automatic extraction of monolingual and bilingual terminology". In: *Proc 15 th COLING*, Kyoto, Japan,1994

- Daille B., "Variations and application-oriented terminology engineering". In *Terminology* 11:1, 2005

- Daille B., et al., "Empirical observation of term variations and principles for their description". In *Terminology* 3(2), 1996

- Dardano, M., "Profilo dell'italiano contemporaneo". In Serianni, L., Trifone, P., (a cura di) *Storia della lingua italiana*, Torino, Einaudi, 1994

- Dardano, M., *La formazione delle parole nell'italiano di oggi*, Roma, Bulzoni Editore, 1978

- De Mauro, T., *Grande Dizionario Italiano dell'Uso*, Torino, UTET, 1999

- Delfitto, D., Paradisi, P., "Prepositionless genitive and N+N compounding in (Old) French and Italian". In *Romance Languages and Linguistic Theory*

2006:Selected papers from 'Going Romance', Amsterdam, Benjamins, 2006

- Dell'Orletta F., "Ensemble system for Part-od-Speech tagging". In *Proceedings of EVALITA 2009 – Evaluation of NLP and Speech Tools for Italian.* Reggio Emilia, Italy, 2009

- Dell'Orletta, F., et al., "Dal testo alla conoscenza e ritorno: estrazione terminologica e annotazione semantica di basi documentali di dominio". In *AIDA Informazioni, Atti del Convegno Nazionale Ass.I.Term "I-TerAnDo"*, Università della Calabria, 5-7-giugno 2008, Roma : AIDA, n. 1-2/2008.

- Diab, M., "Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, PoS tagging, and base phrase chunking". In *2nd International Conference on Arabic Language Resources and Tools*, Cairo, Egypt, 2009

- Diki-Kidiri, M., "Une aproche culturelle de la terminologie". In Terminologie et diversité culturelle, 2000

- Dunning, T., "Accurate Methods for the Statistics of Surprise and Coincidence". In *Computational Linguistics*, 19(1), 1993

- Dury, P., Lervad, S., "La variation synonymique dans la terminologie de l'énergie: approches synchronique et diachronique, deux études de cas", *LSP&Professional Communication* , vol. 8, n. 2, 2008

- El Mahdaouy A., et al., "A Study of Association Measures and their Combination for Arabic MWT Extraction". In *Proceedings 10th International Conference on Terminology and Artificial Intelligence*, Paris, France, 2013.

- Elmgrab, R.A., "Methods of Creating and Introducing New Terms in Arabic: contributions from English-Arabic Translation". In *International Conference on Languages, Literature and Linguistics*, IPEDR vol.26 (2011), Singapore, IACSIT Press

- Faber, P., “The Cognitive Shift in Terminology and Specialized Translation”. In *Monografías de Traducción e Interpretación,* Universitat de València, 2009

- Fahmi, H.H., *AlmrjE fY tEryb AlmSTlHAt AlElmyp wAlfnyp wAlhndsyp*, Il Cairo, AlnhDp AlmSryp, 1961

- Fantinuoli, C. & Zanettin, F., “Creating and using multilingual corpora in translation studies”. In *New directions in corpus-based translation studies*, Berlin:Language Science Press, 2015

- Favretti R. et al., “Words from Bononia Legal Corpus”. In *Text Corpora and Multilingual Lexicography*, John Benjamins, 2007

-Felber, H., Picht, H., *Métodos de terminografía y principios de investigación terminológica,* Madrid, Instituto Miguel de Cervantes, 2006

- Fernández, S., & Kerremans K., “Terminological variation in source texts and translations: A pilot study”. In *Meta: Translators’ Journal*, 56(2), 2011

- Fernández, S. et al., “Multiple motivations in the denomination of concepts: the case of “production area” in the terminology of aquaculture in French and Galician”. In *Terminology Science and Research,* 2009, Visto il 13 Marzo, 2016,https://www.academia.edu/17595787/The_multiple_motivation_in_the_denomination_of_concepts

- Foo. J., *Computational Terminology: Exploring Bilingual and Monolingual Term Extraction*. Licenciate Thesis. Linköping University, Department of Computer and Information Science, NLPLAB - Natural Language Processing Laboratory, 2012

- Frantzi K., Ananiadou S., “The C–value / NC Value domain independent method for multi–word term extraction”. In Journal of Natural Language Processing, 6(3), 1999

- Freixa, J., “Causes of denominative variation in terminology: A typology proposal”. In *Terminology,* 12(1), 2006

- Freixa, J., *La variació terminològica. Analisi de la variació denominativa en textos de diferent grau d'especialització de l'àrea de Medi Ambient*, Tesi di dottorato, Universitat de Barcelona, 2002

- Fung. P. “A Statistical view on Bilingual Lexicon Extraction: From Parallel Corpora to non-Parallel Corpora”. In *Parallel Text Processing:Alignment and Use of Translation Corpora*, Kluwer Academic Publishers, 2000

- Gale, W.A., and Church, W.K. “Identifying word correspondences in parallel texts”. In *Proceedings of the DARPA Speech and Natural Language Workshop* , Pacific Grove, CA, 1991

- Gamper, J., “CATEX– A Project Proposal”. In *Academia*, 14, 10-12, 1998

- Gandin, S., “Linguistica dei corpora e traduzione: definizioni, criteri di compilazione e implicazioni di ricerca dei corpora paralleli”. In *AnnalSS 5*, 2005 (2009), p.134

- Gaudin F.. “La socioterminologie”. In *Langages*, 39e année, n°157, 2005.

- Gotti, M., *I linguaggi settoriali*, Firenze, La Nuova Italia, 1991

- Graff, D., et al., *Standard Arabic Morphological Analyzer (SAMA) Version 3.1*. Linguistic Data Consortium LDC2009E73, 2009

- Grossman, M, Rainer, F. (a cura di), *La formazione delle parole in italiano*, Tübingen: *Niemeyer*, 2004

- Habash, N., Sadat F., "Arabic Preprocessing Schemes for Statistical Machine Translation". In *Proc. of the Human Language Technology Conference of the NAACL*, New York City, NY, 2006.

- Habash, N., “Arabic morphological representations for machine translation”. In *Arabic Computational Morphology: Knowledge- based and Empirical Methods, Text, Speech, and Language Technology*. Kluwer/Springer, 2005

- Habash, N., Rambow, O., “Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In *Proceedings of the Conference of American Association for Computational Linguistics* (ACL 05). Ann Arbor, Michigan, 2005

- Haliday, M. A. K, Hasan, R., *Cohesion in English*. London, Longman, 1976

- Hammouda S, “Small Parallel Corpora in an English-Arabic Translation Classroom: No Need to Reinvent the Wheel in the Era of Globalization”. In Shiyab, S., et al.(a cura di), *Globalization and Aspects of Translation*, UK: Cambridge Scholars Publishing, 2010

- Hamon T., Nazarenko, A. “Detection of synonymy link between terms: Experiment and results”. In Bourigault D., et al (a cura di), *Recent Advances in Computational Terminology*, volume 2 of Natural Language Processing, John Benjamins, 2001

- Harris, Z., *Mathematical structures of language,* New York, Interscience Publishers, 1968

-Irgl, V., “Synonymy in the language of business and economics”. In Laurén, C. and M. Nordman (a cura di) *Special Language. From Human Thinking to Thinking Machines*, Philadelphia: Multilingual Matters LTD, 1989

- Jacquemin, C. “A symbolic and surgical acquisition of terms through variation". In Wermter, S., et al. (a cura di) *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, Heidelberg: Springer, 1996

- Jacquemin, C. et al., “Expansion of multi-word terms for indexing and

retrieval using morphology and syntax". In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, Stroudsburg, USA, ACL, 1997

- Jacquemin, C., "Syntagmatic and paradigmatic representations of term variation" In *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics* (ACL'99)

- Jacquemin, C., Royaute, J., "Retrieving terms and their variants in a lexicalized unification-based framework". In *Proceeding SIGIR '94 Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval,*1994

- Johansson, S., "Reflection on Corpora and their Uses in Cross-linguistic Research" in Zanettin, F., et al.,(a cura di) *Corpora in Translator*, Manchster, St. Jerome: 2003

-Jori, M., "Definizioni e livelli di discorso giuridico". In Scarpelli, U.&Di Lucia, P. (a cura di), *Il linguaggio del diritto*, Milano: LED, 1994

- Justeson J.S., Katz. S.M., "Technical terminology: some linguistic properties and an algorithm for identification in text". In *Natural Language Engineering,* 1:9-27, 1994

- Kageura, K., & Umino, B., "Methods of automatic term recognition". In *Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication*, 3, 1996

- Karin C. Ryding, *A., reference Grammar of Modern Standard Arabic*, Cambridge University Press, 2005

- Kennedy, G., *An Introduction to Corpus Linguistics*, London; New York: Longman, 1998

- Khasarah, M.M, *Elm AlmSTlH wTrA}q wDE AlmSTlHAt fy AlErbyp*, Damasco, Dar El Fikr, 2008

- Koller, W., "The concept of equivalence and the object of translation studies". In *Target* 7, n.2., Amsterdam, John Benjamin, 1995

- Lahbib W., et al. "Arabic -English domain terminology extraction from aligned corpora". In *On the Move to Meaningful Internet Systems: OTM* 2014 Conferences. Lecture Notes in Computer Science, Vol. 8841, Springer, 2014

- Lakoff, G. *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*, Chicago- Londoni, Univerisity of Chicago Press, 1987

- Le An Ha, *Advances in Automatic Terminology Processing: Methodology and application in focus*. PhD Thesis, University of Wolverhampton, UK, 2007

- Lenci, A., et al., *Testo e computer: elementi di linguistica computazionale*, Roma, Carocci editore, 2012

- Li, H., et al., "Comparison of Google Translation with Human Translation". In *Proceedings of the Twenty-Seventh International Florida Artificial Intelligence Research Society Conference,* 2014

- Lin, D.,"Automatic retrieval and clustering of similar words". In *Proceeding COLING '98 Proceedings of the 17th international conference on Computational linguistics - Volume 2, 1998*

- Lyons, J. *Introduction to the theoretical linguistics*, London, Cambridge University Press, I968

- Manning, C., Schütze H. (a cura di) *Foundations of Statistical Natural Language Processing*, Cambridge, Massachusetts, MIT Press, 1999

- Mantovani, D., “Lingua e diritto. Prospettive di ricerca fra sociolinguistica e pragmatica”. In *Il linguaggio giuridico. Prospettive interdisciplinari* (Associazione Italiana Giuristi di Impresa), Milano, Giuffrè, 2008

- Mantovani, D., Pellecchi, L., *La lingua del diritto:formazione, uso, comunicazione*, http://dsg.unipv.it/didattica/insegnamenti/la-lingua-del-diritto-formazione-uso-comunicazione.html, consultato il 1-6-2016

- Manuel, B. *Linguistica dei corpora e linguistica dei corpora italiana. Un’introduzione*, Milano, Qu.A.S.A.R.s.r.l, 2013

- Mayer, F., “Sinonimia ed equivalenza”. In Magris, M., et al., (a cura di) *Manuale di terminologia*, Milano, Hoepli, 2002

- Mcinnes, B.T., *Extending the Log Likelihood Measure to Improve Collocation Identification*, Master's thesis, University of Minnesota, 2004

- Melamed, I. D., “Models of translational equivalence among words”. In *Computational Linguistics,* 26, 2, 2000

- Mitkov, R., et al., “A new, fully automatic version of Mitkov's knowledge poor pronoun resolution method”. In *Proceedings of CICLing,* 2002, Mexico City, Mexico

- Montiel-Ponsoda E., et al., “Multilingual variation in the context of linked data”. In *Proceedings of the 10th International Conference on Terminology and Artificial Intelligence,* 2013

- Moore, R. C., “Towards a simple and accurate statistical approach to learning translation relationships among words”. In *Proceedings of the ACL workshop on data-driven machine translation*, Toulouse, France, 2001

- Mortara Garavelli B., “Persistenza del latino nell'uso giuridico odierno”. In *L'Accademia della Crusca per Giovanni Nencioni*, Firenze, Le Lettere, 2002

- Mortara Garavelli, B., *Le parole e la giustizia: Divagazioni grammaticali e retoriche su testi giuridici italiani,* Torino, Giulio Einaudi Editore, 2001

- Nandor, B., "Sostantivi composti nell'italiano contemporaneo". In *Lingua Nostra*, XXXIX / 4, 1978

- Neveu F., "Un aspect de l'apport des corpus à la terminologie linguistique: l'alignement". In *Mots, Termes, et Contextes,* actes des journées scientifiques du réseau Lexicologie, Terminologie, Tradution de l'Agence Universitaire de la Francophonie, Paris, Editions des Archives Contemporaines, 2006

- Och, F., *Statistical Machine Translation: From Single-Word Models to Alignment Templates*. PhD thesis, RWTH Aachen, Germany, 2002

- Och, F., *An Efficient* Method for *Determining Bilingual Word Classes.* In EACL '99:Ninth Conf. Of the Europ.Chapter of the assocication for computational linguistics, Bergen, Norway, 1999

- Olohan, M., "Introducing Corpora". In *Translation Studies*, Routledge, London & New York, 2004

- Ouhalla, J., Shlonsky, U. *Themes in Arabic and Hebrew Syntax*, Kluwer Academic Pub, 2002

- Pasha, A., et al., "MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic". In *Language Resources and Evaluation Conference(LREC)*, Reykjavik, 2014.

- Pazienza, M.T., et al.,: "Terminology extraction: an analysis of linguistic and statistical approaches". In *Knowledge Mining*, Springer, 2005

- Pelletier, J., La variation terminologique : un modèle à trois composantes, Philosophiæ doctor (Ph.D.), Université Laval, 2012

- Peruzzo, K., *Terminological Equivalence and Variation in the EU Multi-level Jurisdiction: A Case Study on Victims of Crime*. Doctoral thesis in Interpreting and Translation Studies, IUSLIT, University of Trieste, 2013

- Peter F. et al. "A statistical approach to machine translation". In *Computational Linguistics* , 16(2):79–85, 1990.

- Peter F. et al.,"The mathematics of statistical machine translation: Parameter estimation". In *Computational Linguistics*, 19(2):263–311, June 1993

-Picchi E. et al., "Risorse monolingui e multilingui. Corpus bilingue italiano-arabo". In *Linguistica computazionale*, XVIII/XIX, 1999, Pisa

- Ralli, N., "Terminografia e comparazione giuridica: metodo, applicazioni e problematiche chiave". In *InTRAlinea, online Translation Journal*, 2010, http://www.intralinea.org/specials/article/1727, cliccato il 30-4-2015

- Reguigui, A., *La Créativité Lexicale en Terminologie Arabe*, Université Laurentienne, Ontario, 1994

- Ricca, D., "Al limite tra sintassi e morfologia: i composti aggettivali V-N nell'italiano contemporaneo". In Grossmann, M., Thornton, A.M., (a cura di), *La formazione delle parole. Atti del XXXVII congresso della Società di Linguistica Italiana*, Roma, Bulzoni, 2005

- Rogers, M., "Synonymy and equivalence in special-language texts. A Case Study in German and English Texts on Genetic Engineering". In Trosborg, A., (a cura di) *Text Typology and Translation,* Amsterdam, John Benjamins, 1997

- Sabatini, F., "Analisi del linguaggio giuridico. Il testo normativo in una tipologia generale dei testi". In *Corso di studi superiori legislativi 1988-1989*, a cura di M. D'Antonio, Padova, CEDAM, 1990

- Sacco, R., "Lingua e diritto". In *Ars Interpretandi* 5, 2000

- Sager, J., *A Practical Course in Terminology Processing*. Amsterdam: John Benjamins Publishing Company, 1990

- Sagri, M., Tiscornia, D., "Le peculiarità del linguaggio giuridico. Problemi e prospettive nel contesto multilingue europeo". In *MediAzioni* 7, 2009, http://mediazioni.sitlec.unibo.it

- Samy, D., et al., "Building a Multilingual Parallel Corpus Arabic-Spanish-English". In *Proceedings of International Conference on Language Resources and Evaluation LREC-06*, Genoa, Italy, 2006

- Sandrini, P., "Comparative Analysis of Legal Terms: Equivalence Revisited". In *TKE* '96. Frankfurt: Indeks

- Saussure, F., *Course in General Linguistics*. Translated by Roy Harris, London: Duckworth, 1983

- Scalise, S., Bisetto, A., "Compounding: Morphology and/or syntax?". In Mereu, L. (a cura di) *Boundaries of morphology and syntax*, Amsterdam/Philadelphia:John Bejamins,1999

- Scalise, S., Bisetto, A., *La struttura delle parole*, Bologna, Il Mulino, 2008

- Scarpa, F. *La traduzione specializzata - Lingue speciali e mediazione linguistica*, Milano: Ulrico Hoepli, 2001

- Scarpelli, U., "Semantica giuridica". In *Novissimo digesto italiano*, vol.XVI, Torino, Utet, 1969

- Schrader. B. *Exploiting Linguistic and Statistical Knowledge in a Text Alignment System*. PhD thesis, Universität Osnabrück, 2007

- Schütze, H., "Dimensions of meaning". In *Proceedings of the ACM/IEEE* conference on Supercomputing, USA, 1992

- Serianni, L. *Italiani scritti,* Bologna, Il Mulino, 2003

- Shannon, C., "A mathematical theory of communication". In *Bel System Technical Journal*, 27:379 423, 1948.

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

- Slodzian M., "L'émergence d'une terminologie textuelle et le retour du sens". In *Le sens en terminologie*, Lyon, Presses Universitaires de Lyon, 2000

- Slodzian, M., "La terminologie, historique et orientations". In Harzallah, M., et al., (a cura di), *Actes de la semaine de la Connaissance* (SDC'06), Conférence Invitée, Nantes, 28-30 Juin 2006.

- Smadja, F., "Retrieving collocations from text: Xtract", in *Computational Linguistics*, Cambridge, MIT Press, 19(1), 1993.

- Temmerman, R., "Questioning the univocity ideal. The difference between sociocognitive Terminology and traditional Terminology". In *Hermes. Journal of Linguistics* 18, 1997

- Teserra, S. A., *Bilingual word and chunk alignment: a hybrid system for Amharic and English*, tesi di dottorato, Bielefeld University, 2007

- Tiedemann, J., "Combining clues for word alignment". In *Proceedings of the 10th Conference of the European Chapter of the ACL (EACL)*, Budapest, Hungary, 2003

- Tiedemann, J., *Recycling translations: Extraction of lexical data from parallel corpora and their application in natural language processing corpora*, tesi di dottorato, Acta Universitatis Upsaliensis, 2003, consultato il 20/02/2016: http://stp.lingfil.uu.se/~joerg/phd/html/

-Tiscornia, D., Sagri, M, T., "Legal Concepts and Multilingual Contexts in Digital Information". In *Beijing Law Review*, Vol.3 No.3, 2012

- Tufiş, D., Barbu, A.M."Lexical token alignment: experiments, results and applications". In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002)*, Las Palmas, Spain, 2002

- Ullmann, S. *The Principles of Semantics*, Glasgow: Jackson and Oxford: Blackwell, 1957

- Van Der Plas L., Tiedemann J., "Finding medical term variations using parallel corpora and distributional similarity". In: *Proceedings of the Coling workshop on ontologies and lexical resources,* Beijing, Cina, 2010.

- Visconti, J., "Prestiti e calchi: dove va la lingua giuridica italiana", in Bambi, F., Pozzo, B., (a cura di), *Dove va l'italiano giuridico*, Firenze, Accademia della Crusca, 2012

- Vogel, S., et al., "HMM-based word alignment in statistical translation". In *Proceedings of the 16th International Confernece on Computational Linguistics*, Copenhagen, Denmark, 1996

- Vulić, I., *Term Alignment. State of the Art Overview*, Katholieke Universiteit Leuven, 2010

- Wharf, B., *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*, Cambridge, MIT Press, 1956

- Wong, W., et al., "Determination of Unithood and Termhood for Term Recognition". In *Handbook of Research on Text and Web Mining Technilogies*, US, IGI Global, 2009

- Wu, H., Ming Zhou, M.,"Optimizing synonym extraction using monolingual and bilingual resources. In *Proceedings of the second international workshop on Paraphrasing*, PA, USA, 2003

- Wüster, E., *Introduction to the General Theory of Terminology and Terminological Lexicography*, Vienna, Springer, 1979

- Yoshikane, F., et al., "Detecting Japanese Term Variation in Textual Corpus". In *Proceedings 4th International Workshop on Information Retrieval with Asian Languages (IRAL'99*), Taipei, Taiwan, Academia Sinica

- Zampolli A., "Introduzione". In Ridolfi P., Piraino, R. (a cura di), *Trattamento automatico della lingua nella Società dell'informazione*. *La Comunicazione*, XLVI, numero unico (1,2,3,4), Roma,1997

- Zbib, R., Soudi, A., "Introduction: Challenges for Arabic Machine Translation". In Soudi, A. et al (a cura di), *Challenges for Arabic Machine Translation*, John Benjamins Publishing Company, 2012

- Zhang, C. et al., "Automatic keyword extraction from documents using conditional random fields". In *Journal of computational and information systems* 4:3, 2008

- Zotti, P., "Costruire un corpus parallelo Giapponese-Italiano. Metodologie di compilazione e applicazioni". In Casari, M., Scrolavezza, P. (a cura di), *Giappone, storie plurali*, Bologna, I libri di Emil-Odoya Edizioni, 2013

- Zanettin, F., *Translation-driven corpora: Corpus resources for descriptive and applied translation studies*. Manchester: St. Jerome Publishing, 2012

- Zuffi S., "The nominal composition in Italian. Topics in generative morphology". in *Journal of Italian Linguistics* 2,1981

Università
Ca'Foscari
Venezia

DEPOSITO ELETTRONICO DELLA TESI DI DOTTORATO
DICHIARAZIONE SOSTITUTIVA DELL'ATTO DI NOTORIETA'
(Art. 47 D.P.R. 445 del 28/12/2000 e relative modifiche)

Io sottoscritto FATHI HASSAN AHMED FAWI nato Kena (Egitto) il 18/12/1982 residente a MESTRE (VE) in VIA BEMBO n.40
Matricola (se posseduta) 955994 Autore della tesi di dottorato dal titolo:
LE VARIAZIONI TERMINOLOGICHE IN UN CORPUS GIURIDICO PARALLELO ITALIANO-ARABO: STUDIO LINGUISTICO -COMPUTAZIONALE
Dottorato di ricerca in SCIENZE DEL LINGUAGGIO
Ciclo 28°
Anno di conseguimento del titolo 2017

DICHIARO

di essere a conoscenza:
1) del fatto che in caso di dichiarazioni mendaci, oltre alle sanzioni previste dal codice penale e dalle Leggi speciali per l'ipotesi di falsità in atti ed uso di atti falsi, decado fin dall'inizio e senza necessità di nessuna formalità dai benefici conseguenti al provvedimento emanato sulla base di tali dichiarazioni;
2) dell'obbligo per l'Università di provvedere, per via telematica, al deposito di legge delle tesi di dottorato presso le Biblioteche Nazionali Centrali di Roma e di Firenze al fine di assicurarne la conservazione e la consultabilità da parte di terzi;
3) che l'Università si riserva i diritti di riproduzione per scopi didattici, con citazione della fonte;
4) del fatto che il testo integrale della tesi di dottorato di cui alla presente dichiarazione viene archiviato e reso consultabile via internet attraverso l'Archivio Istituzionale ad Accesso Aperto dell'Università Ca'Foscari, oltre che attraverso i cataloghi delle Biblioteche Nazionali Centrali di Roma e Firenze;
5) del fatto che, ai sensi e per gli effetti di cui al D.Lgs. n. 196/2003, i dati personali raccolti saranno trattati, anche con strumenti informatici, esclusivamente nell'ambito del procedimento per il quale la presentazione viene resa;
6) del fatto che la copia della tesi in formato elettronico depositato nell'Archivio Istituzionale ad Accesso Aperto è del tutto corrispondente alla tesi in formato cartaceo, controfirmata dal tutor, consegnata presso la segreteria didattica del dipartimento di riferimento del corso di dottorato ai fini del deposito presso l'Archivio di Ateneo, e che di conseguenza va esclusa qualsiasi responsabilità dell'Ateneo stesso per quanto riguarda eventuali errori, imprecisioni o omissioni nei contenuti della tesi;
7) del fatto che la copia consegnata in formato cartaceo, controfirmata dal tutor, depositata nell'Archivio di Ateneo, è l'unica alla quale farà riferimento l'Università per rilasciare, a richiesta, la dichiarazione di conformità di eventuali copie.

Data _____ Firma _____

Mod. TD-Lib-09-a 1

AUTORIZZO

- l'Università a riprodurre ai fini dell'immissione in rete e a comunicare al pubblico tramite servizio on line entro l'Archivio Istituzionale ad Accesso Aperto il testo integrale della tesi depositata;
- l'Università a consentire:
- la riproduzione a fini personali e di ricerca, escludendo ogni utilizzo di carattere commerciale;
- la citazione purché completa di tutti i dati bibliografici (nome e cognome dell'autore, titolo della tesi, relatore e correlatore, l'università, l'anno accademico e il numero delle pagine citate).

DICHIARO

1) che il contenuto e l'organizzazione della tesi è opera originale da me realizzata e non infrange in alcun modo il diritto d'autore né gli obblighi connessi alla salvaguardia di diritti morali od economici di altri autori o di altri aventi diritto, sia per testi, immagini, foto, tabelle, o altre parti di cui la tesi è composta, né compromette in alcun modo i diritti di terzi relativi alla sicurezza dei dati personali;
2) che la tesi di dottorato non è il risultato di attività rientranti nella normativa sulla proprietà industriale, non è stata prodotta nell'ambito di progetti finanziati da soggetti pubblici o privati con vincoli alla divulgazione dei risultati, non è oggetto di eventuali registrazione di tipo brevettuale o di tutela;
3) che pertanto l'Università è in ogni caso esente da responsabilità di qualsivoglia natura civile, amministrativa o penale e sarà tenuta indenne a qualsiasi richiesta o rivendicazione da parte di terzi.

A tal fine:
- dichiaro di aver autoarchiviato la copia integrale della tesi in formato elettronico nell'Archivio Istituzionale ad Accesso Aperto dell'Università Ca' Foscari;
- consegno la copia integrale della tesi in formato cartaceo presso la segreteria didattica del dipartimento di riferimento del corso di dottorato ai fini del deposito presso l'Archivio di Ateneo.

Data _____ **Firma** _____

La presente dichiarazione è sottoscritta dall'interessato in presenza del dipendente addetto, ovvero sottoscritta e inviata, unitamente a copia fotostatica non autenticata di un documento di identità del dichiarante, all'ufficio competente via fax, ovvero tramite un incaricato, oppure a mezzo posta

Firma del dipendente addetto .....

Ai sensi dell'art. 13 del D.Lgs. n. 196/03 si informa che il titolare del trattamento dei dati forniti è l'Università Ca'Foscari - Venezia.
I dati sono acquisiti e trattati esclusivamente per l'espletamento delle finalità istituzionali d'Ateneo; l'eventuale rifiuto di fornire i propri dati personali potrebbe comportare il mancato espletamento degli adempimenti necessari e delle procedure amministrative di gestione delle carriere studenti. Sono comunque riconosciuti i diritti di cui all'art. 7 D. Lgs. n. 196/03.

Mod. TD-Lib-09-a 2

Estratto per riassunto della tesi di dottorato

Studente: FATHI HASSAN AHMED FAWI matricola: 955994
Dottorato: SCIENZE DEL LINGUAGGIO
Ciclo: 28°
Titolo della tesi: LE VARIAZIONI TERMINOLOGICHE IN UN CORPUS GIURIDICO PARALLELO ITALIANO-ARABO: STUDIO LINGUISTICO -COMPUTAZIONALE

Abstract:
La presente tesi si pone l'obiettivo di studiare le variazioni terminologiche in un corpus giuridico parallelo italiano-arabo, adottando un approccio linguistico-computazionale. La tesi parte dall'assunto che anche i linguaggi specialistici presentano delle variazioni a livello lessicale come è il caso della lingua comune, contrastando quindi con la teoria generale della terminologia basata sul principio di monosemia e univocità.
Nel lavoro la componente linguistica si integra con i metodi statistici: mentre la parte linguistica si evidenzia nei capitoli riguardanti la variazione nei discorsi specializzati, i procedimenti di formazione delle parole in italiano e in arabo e l'analisi delle variazioni estratte dal corpus parallelo; le misure statistiche vengono adoperate, oltre che nella creazione e nell'annotazione del corpus parallelo, nell'estrazione, sia monolingue che bilingue, dei termini dal corpus e nell'individuazione delle variazioni terminologiche.

This thesis aims to study the terminological variations in an Italian-Arabic parallel corpus of legal texts, adopting a computational linguistic approach. The thesis starts from the assumption that even the specialized languages can have variations on the lexical level, contrasting with the general theory of terminology based on the principle of monosemy and univocity.
In this work the linguistic analysis is integrated with the statistical methods: while the linguistic component is evident in the chapters regarding the variation in specialized discourses and the formation of words in Italian and Arabic, the statistical measures are used to create and annotate the parallel corpus, to extract the monolingual and bilingual terms from the corpus and finally to identify the terminological variations.

Firma dello studente _____