Maria Pia di Buono

RAPPRESENTAZIONE FORMALE DEGLI ASPETTI MORFO-SINTATTICI IN APPLICAZIONI MULTILINGUE: IL CASO DEI CLITICI NELLE ESPRESSIONI POLIREMATICHE VERBALI ITALIANE

Abstract

Questo articolo presenta i risultati della rappresentazione formale nel formato OntoLex-*lemon* delle proprietà morfo-sintattiche dei clitici nelle espressioni polirematiche verbali dell'italiano. A partire dall'analisi degli aspetti sintattici e morfologici e di restrizioni combinatorie che caratterizzano tali particelle, viene proposto un set di regole in grado di rappresentare i diversi livelli informativi al fine di sviluppare una risorsa linguistica interoperabile.

Parole chiave: espressioni polirematiche verbali, clitici, LLOD, rappresentazione formale, morfologia

This paper presents the results of an OntoLex-*lemon* formal representation of morpho-syntactic aspects of clitics in Italian verbal multiword expressions. Starting from the analysis of morphological and syntactic features and combinatorial behaviours of those particles, we propose a set of rules suitable for representing different levels of linguistic information in order to develop an interoperable linguistic resource.

Keywords: Verbal multiword expressions, clitics, LLOD, formal representation, morphology

1. Introduzione

La rappresentazione formale dei dati linguistici in un formato matematico o interpretabile dalle macchine costituisce uno snodo di ricerca

Maria Pia di Buono, Dipartimento di Studi Letterari, Linguistici e Comparati, Università di Napoli L'Orientale, mpdibuono@unior.it.

fondamentale nello sviluppo di sistemi per il Trattamento Automatico del Linguaggio (TAL) e per applicazioni di Intelligenza Artificiale (IA). Infatti, una rappresentazione coerente e consistente del senso dei dati linguistici può contribuire direttamente a diversi task nei due campi, ad esempio per i task di Word Sense Disambiguation (Navigli, 2009), similarità semantica (Budanitsky and Hirst, 2006; Turney and Pantel, 2010; Pilehvar *et al.*, 2013), allineamento di risorse lessicali (Navigli and Ponzetto, 2012; Pilehvar and Navigli, 2014; Niemann and Gurevych, 2011), di creazione di inventari di sensi (Navigli, 2006; Snow *et al.*, 2007), di sostituzioni lessicali (McCarthy and Navigli, 2009), di semantic priming (Neely *et al.*, 1989). Una tale rappresentazione a livello di sensi di parole o di sensi di unità linguistiche minori, come morfemi, fonemi, etc., può essere inoltre direttamente estesa nello sviluppo di ambienti e applicazioni multilingui con diverse finalità.

Per certi scopi, come ad esempio nel caso in cui la risorsa linguistica sia utilizzata da traduttori umani o apprendenti L2, una rappresentazione puramente lessicale con l'entrata, sia essa una parola semplice o un'espressione polirematica, la sua corrispondenza nella seconda lingua ed eventualmente un esempio d'uso è sufficiente. Nei casi in cui l'obiettivo sia quello di riutilizzare la stessa risorsa all'interno di applicazioni TAL e IA, al fine di assicurare che l'entrata lessicale sia trattata automaticamente in maniera adeguata, ci sarà bisogno probabilmente di aggiungere altre informazioni, come le proprietà morfosintattiche e il suo specifico comportamento. Inoltre, le entrate lessicali e, in particolar modo, le espressioni polirematiche presentano differenti caratteristiche intrinseche e le informazioni necessarie per ogni particolare entrata possono variare sulla base della tipologia considerata e dell'uso finale delle risorse linguistiche (Escartín *et al.*, 2013).

Per tali motivi, in anni recenti, alle finalità di interoperabilità e riusabilità delle risorse sviluppate, si è aggiunta la necessità di garantire anche la rappresentatività formale al fine di favorire la connessione delle risorse con altre risorse esistenti e integrare informazioni da basi di conoscenza esterne, creando in questo modo una rete interconnessa di dati linguistici. La rilevanza assunta dalla rappresentazione formale è evidente nel

proliferare di studi scientifici, progetti e iniziative, a livello europeo e non solo, che supportano lo sviluppo di modelli per la rappresentazione dei dati linguistici e in particolar modo delle espressioni polirematiche.

Le attuali metodologie per la modellazione e la rappresentazione formale dei dati linguistici sono ampiamente basate sull'utilizzo di ontologie, in particolar modo per favorire l'integrazione di tali informazioni nel Web semantico e nelle emergenti tecnologie per il linguaggio basate sulla semantica. Gli standard ontologici disponibili per la modellazione dei dati in diverse varietà di forma, come ad esempio il Web Ontology Language (OWL) (McGuinness and Harmelen, 2004), risultano poco flessibili per la formalizzazione dei dati linguistici. La scarsa flessibilità di tali modelli riduce la possibilità di rappresentare le informazioni linguistiche, limitando la descrizione dettagliata delle entità lessicali. In particolare, la gestione di una rappresentazione efficace delle espressioni polirematiche risulta complessa, soprattutto quando rivolta a fini di interoperabilità e riusabilità delle risorse linguistiche, ma è essenziale per assicurare la loro piena integrazione nelle applicazioni di TAL, nei loro workflow e nelle loro infrastrutture (Escartín *et al.*, 2013). In una prospettiva monolingue e multilingue, una rappresentazione formale efficace dovrà tenere adeguatamente in conto le proprietà semantiche e morfo-sintattiche dell'espressione - intesa come unità linguistica unica - e dei suoi componenti, nonché la struttura interna e le dipendenze, la variazione sintattica, e, potenzialmente, anche le varietà linguistiche regionali (Escartín et al., 2013).

Il contributo di questo lavoro mira a proporre soluzioni di formalizzazione compatibili con le attuali tecnologie semantiche del linguaggio per garantire una corretta rappresentazione formale delle espressioni polirematiche verbali dell'italiano che presentano elementi clitici. Questa scelta è motivata dal fatto che, come sottolineato da Berretta (1984) nel suo studio sulle sequenze di apprendimento per l'italiano come L2, i pronomi clitici costituiscono "un sottosistema morfologico parecchio articolato, marcato, caratteristico di alcune lingue romanze, comprendente forme atone e toniche, con collocazione sintattica variabile rispetto alla posizione verbale".

Le soluzioni di modellazione formale adottate per la rappresentazione di queste informazioni confluiscono in una risorsa linguistica il cui sviluppo contribuisce ad arricchire le risorse linguistiche per l'italiano e allo stesso tempo assicura che lo sviluppo dei modelli, attualmente in corso, risponda in maniera adeguata alle caratteristiche morfologiche di questi fenomeni.

L'articolo è strutturato come segue: nella Sezione 2 si introduce l'importanza della rappresentazione formale dei dati linguistici nello sviluppo di risorse interoperabili, successivamente vengono discusse le caratteristiche dei clitici nelle espressioni polirematiche verbali italiane a partire dagli studi già sviluppati in tale ambito (Sezione 3). Nella Sezione 4 si presenta il modello scelto per la formalizzazione e il set di regole morfologiche, fonologiche e sintattiche proposte per la rappresentazione degli aspetti morfo-sintattici dei clitici. Infine, le conclusioni di tale ricerca vengono discusse nella Sezione 5.

2. Rappresentazione formale dei dati linguistici[1]

In tutte le aree della linguistica empirica, della filologia computazionale e in applicazioni per il TAL, l'annotazione linguistica e l'utilizzo di marcature (*markup*), basate su linguaggi descrittivi e, generalmente, realizzate in formato Extensible Markup Language (XML)[2], rappresentano un elemento centrale di analisi.

Nonostante il formato XML e gli altri linguaggi descrittivi siano particolarmente diffusi, in quanto offrono la possibilità di realizzare marcature che siano comprensibili sia ad utenti umani che alle macchine, l'impossibilità di garantire l'interazione delle strutture in maniera gerarchica, in altre parole di evitare la sovrapposizione delle marcature[3], rappresenta un problema per tali formati. Per superare questi limiti dei

[1] Questa sezione è parzialmente basata sulla pagina Wikipedia dei Linguistic Linked Open Data (https://en.wikipedia.org/wiki/Linguistic_Linked_Open_Data), che l'autrice ha contribuito a redigere e tradurre nella versione in lingua italiana.

[2] https://en.wikipedia.org/wiki/XML

[3] Questo problema è conosciuto con il termine di *overlapping markup* o *concurrent markup*.

linguaggi descrittivi, a partire dalla fine degli anni '90, sono stati sviluppati modelli di dati basati sui grafi (Bird and Liberman, 1998), che consentissero l'interconnessione di molteplici file XML *(standoff XML*[4]*)*. Tuttavia, i modelli basati sui grafi non sempre sono adeguatamente supportati dalla tecnologia XML disponibile (Eckart, 2008) e il progresso in tale campo ha subìto dei rallentamenti anche a causa della scarsa interoperabilità, imputabile soprattutto alle differenze nei vocabolari e negli schemi di annotazione usati per differenti risorse e strumenti.

Al fine di ovviare ai limiti degli esistenti modelli descrittivi e connettere le risorse linguistiche e le banche dati di ontologie/terminologie, facilitando il ri-uso di vocabolari condivisi e l'interpretazione degli stessi rispetto ad una base comune, è stata adottata una soluzione basata sull'uso dei Linked Data[5], formalizzati nel formato Resource Description Framework (RDF).

L'applicazione dei principi dei Linked Data per la rappresentazione delle informazioni relative ai dati linguistici ha dato vita ai Linguistic Linked Open Data (LLOD)[6], espressione con la quale ci si riferisce sia al metodo che alla comunità che si occupa di creare, condividere e (ri) utilizzare risorse linguistiche che applicano tali principi.

Lo sviluppo e la pubblicazione di dati linguistici e di risorse per l'elaborazione del linguaggio naturale secondo i principi di modellazione propri dei LLOD, identificati da Chiarcos *et al.* (2013), comportano una serie di benefici che interessano diversi aspetti: la rappresentazione, l'interoperabilità, la federazione, la creazione di un ecosistema di dati linguistici provenienti da diverse sorgenti, la semantica, la dinamicità. L'applicazione dei principi e della metodologia dei LLOD consente, inoltre, di superare i

[4] ISO 24612:2012. "Language resource management - Linguistic annotation framework (LAF)". ISO. Visitato il 2020-01-25.

[5] https://en.wikipedia.org/wiki/Linked_data#Linked_open_data

[6] La ''Linguistic Linked Open Data Cloud'' (http://linguistic-lod.org/llod-cloud) sviluppata e sostenuta dal gruppo di lavoro Open Linguistics Working Group (OWLG) della Open Knowledge Foundation (in italiano Fondazione per la conoscenza aperta), ha rappresentato, sin dalla nascita, il centro focale delle attività di diversi gruppi delle comunità afferenti al World Wide Web Consortium (W3C), progetti di ricerca, e sforzi per la creazione di risorse e infrastrutture dedicate al sostenere questa iniziativa.

problemi relativi alle applicazioni multilingue, incluse l'interconnessione di risorse lessicali eterogenee, come ad esempio WordNet[7] e Wikipedia.

Con l'obiettivo di migliorare le possibilità di modellazione formale su base semantica dei dati linguistici e di sfruttare i vantaggi offerti dall'applicazione dei principi LLOD, diversi modelli sono stati rilasciati nel corso degli anni. Il superamento dei limiti dei primi modelli di formalizzazione avviene grazie allo sviluppo del modello *lemon* (McCrae et al., 2012), uno dei primi sistemi formali orientati alla rappresentazione dei dati linguistici. Il modello *lemon* è basato sulla creazione di un lessico in grado di descrivere con un dettaglio maggiore le informazioni delle realizzazioni linguistiche di un concetto. In questa ricerca, si è scelto di applicare il *Lexicon model for ontologies*[8], OntoLex-*lemon*, un modello di rappresentazione per la formalizzazione dei dati linguistici riconosciuto dal World Wide Web Consortium[9] (W3C), nell'ambito dello sviluppo di standard open che assicurino la crescita del Web a lungo termine. Tale scelta è giustificata dall'ampia diffusione del modello, dalla comunità di sviluppo e supporto particolarmente attiva e, non da ultimo, dalle capacità rappresentative che il modello offre.

OntoLex-*lemon* e la rappresentazione morfologica. Il modello OntoLex-*lemon* è stato oggetto di sviluppo per molti anni ed è stato, inizialmente, derivato dall'integrazione di tre modelli pre-esistenti: LingInfo (Buitelaar *et al.* 2006), LexOnto (Cimiano *et al.*, 2007), LIR (Montiel-Ponsoda *et al.*, 2011).

Lo sviluppo di un modello basato su ontologie per la rappresentazione delle informazioni lessicali costituisce, infatti, uno dei risultati del lavoro dell'OntoLex Community Group[10], fondato nel 2011 (McCrae *et al.*, 2017). I modelli e i relativi vocabolari sono stati sviluppati in maniera induttiva a partire dalla raccolta di casi d'uso e dalla successiva analisi di questi al fine di derivare una serie di requisiti fondamentali per questo tipo di rappresentazione formale.

[7] WordNet è una base di dati lessicale sviluppata inizialmente per la lingua inglese e successivamente ampliato ad altre lingue. https://wordnet.princeton.edu/

[8] https://www.w3.org/2016/05/ontolex/

[9] https://www.w3.org/

[10] https://www.w3.org/community/ontolex/

Lo sviluppo del primo modello è avvenuto in due fasi: i) sviluppo del *core module* (modulo principale) al fine di introdurre gli elementi descrittivi di base che sono stati individuati, a partire dai casi studio analizzati, come esaustivi di tutte le possibili applicazioni del modello; ii) sviluppo di moduli aggiuntivi per la rappresentazione di altri aspetti linguistici specifici. Il *core module* e i moduli aggiuntivi sono stati successivamente unificati e documentati in un rapporto tecnico (Cimiano *et al.*, 2016) pubblicato dal W3C insieme ai file del modello tecnico in OWL.

Il modello principale di OntoLex-*lemon* è strutturato a partire da un'entrata lessicale rappresentata come una `ontolex:LexicalEntry` (Figura 1) che include le sottoclassi di parola (semplice), espressioni polirematiche e affissi. Come descritto da McCrae et al. (2017), una `ontolex:LexicalEntry` rappresenta una singola entrata lessicale direttamente connessa con diverse classi del modello adatte a rappresentare le informazioni lessicali su diversi livelli: `ontolex:Form`,utilizzata per modellare tutte le espressioni morfologiche di un'entrata lessicale; `ontolex:LexicalSense` per il mapping tra un'entrata lessicale e un concetto e, infine, `ontolex:LexicalConcept` per specificare un significato senza dover ricorre ad un'ontologia esterna.

Il modello supporta anche la formalizzazione di espressioni polirematiche e la definizione di regole d'uso per una particolare entrata lessicale che esprime un dato concetto, consentendo di annotare il senso di un lemma così da specificare il suo ruolo nel mapping tra entrate lessicali e concetti.

```
@prefix ontolex:
<http://www.w3.org/ns/lemon/ontolex#> .
@prefix : <#>.

:lex_child a ontolex:LexicalEntry ;
  ontolex:lexicalForm :form_child_singular,
:form_child_plural .

:form_child_singular a ontolex:Form ;
ontolex:writtenRep “child”@en .
```

```
:form_child_plural a ontolex:Form ;
  ontolex:writtenRep "children"@en .
```

Figura 1. *Esempio di una lexical entry che presenta due form*[11]

Al modulo centrale di OntoLex-*lemon* si affianca un modello per la rappresentazione delle informazioni morfologiche (Morph module) per il cui sviluppo sono stati presi in considerazione diversi aspetti che hanno guidato le scelte di modellazione: (i) scopo e copertura, (ii) consistenza, (iii) ambiguità terminologica (Klimek *et al.*, 2019).

Per quanto riguarda lo scopo generale e la copertura dei diversi aspetti morfologici, il modello intende consentire: (i) la rappresentazione degli elementi che sono coinvolti nei meccanismi di composizione/derivazione delle entrate lessicali e delle forme di parola (*word form*), (ii) la rappresentazione dei pattern e delle regole di formazione e flessione delle entrate lessicali e delle forme di parole.

Al fine di garantire la consistenza della rappresentazione rispetto ad altri modelli già esistenti, Klimek *et al.* (2019) hanno preso in considerazione gli elementi di OntoLex-*lemon*, classi e proprietà, e quelli provenienti da vocabolari esterni. Tale scelta è motivata dalla necessità di riutilizzare elementi di modellazione precedentemente sviluppati per ridurre l'introduzione di nuove classi e proprietà, i cui scopi potrebbero essere in sovrapposizione o contrasto con gli elementi di altri modelli. Tuttavia, il riutilizzo delle classi e proprietà vincola lo sviluppo di un modello, in quanto non è possibile ridefinire la funzione rappresentativa degli elementi esistenti, limitando, di fatto, la possibilità di modifiche soltanto alle restrizioni di dominio (*domain*) ed estensione (*range*) delle proprietà.

Infine, un altro elemento di valutazione è stata la definizione della terminologia da impiegare per evitare ambiguità nelle definizioni, in termini di etichette adottate per le classi e le proprietà nel nuovo modello.

[11] Esempio tratto da https://www.w3.org/community/ontolex/wiki/Final_Model_Specification

3. I clitici nelle espressioni polirematiche verbali italiane

I clitici in italiano rappresentano un oggetto di analisi largamente approfondito da molti studiosi[12], in quanto presentano aspetti sintattici e morfologici e di restrizioni combinatorie piuttosto variegati. Inoltre, la categoria delle costruzioni verbali che presentano clitici è altamente produttiva in italiano e include costruzioni verbali con diverse strutture: (i) riflessive; (ii) reciproche; (iii) pronominali; a cui si aggiungono gli impieghi di natura idiomatica e quasi idiomatica (Masini, 2012). Data la complessità del fenomeno, numerose ricerche si sono concentrate sull'analisi di diversi aspetti di queste occorrenze. Ai fini di questa analisi, seguendo la classificazione proposta da Savary *et al.* (2015) e da Monti & di Buono (2019), vengono presi in considerazione i verbi intrinsecamente riflessivi (Inherently Reflexive Verbs - IRV), inerentemente clitici[13] (Language Specific - Inherently Clitic Verbs - LS.ICV) e le occorrenze idiomatiche e quasi idiomatiche (Verbal Idioms - VID)[14].

I verbi intrinsecamente riflessivi, denominati anche riflessivi inerenti (Benincà et al. 1988), sono verbi intransitivi che presentano uno o più clitici e nei tempi composti utilizzano l'ausiliare *essere* in quanto inaccusativi. Il pronome riflessivo presente in queste costruzioni non è un argomento (De Alencar & Kelling, 2005) e, quindi, non fa parte della struttura argomentale del verbo stesso, viene per questo motivo considerato un pronome espletivo, o riflessivo desemantizzato (Cordin, 1988).

Due sono i casi in cui è possibile distinguere i verbi intrinsecamente riflessivi: (a) quando il verbo non occorre senza il clitico non riflessivo, come nel caso di *suicidarsi*; (b) quando il verbo con il clitico e quello sen-

[12] A titolo esemplificativo, si ricordano tra gli altri i lavori di Cennamo (1993), Cinque (1988), Lo Cascio (1970), Simone (1983)

[13] Procomplementari secondo la definizione di De Mauro nel GRADIT (1999-2000).

[14] I nomi delle categorie verbali sono quelle utilizzate nell'ambito della COST Action PARSEME (Savary *et al.*, 2015) e nell'annotazione del corpus PARSEME-It VMWE (Monti & di Buono, 2019). Per motivi di sintesi, non si riportano le descrizioni per ciascuna delle categorie di VMWE individuate nel corpus, per le quali si rimanda alla letteratura di riferimento.

za clitico presentano due significati diversi, come *fare* e *farsi*, o realizzano una diversa sottocategorizzazione o selezione (Chomsky, 1965) in base alle varie classi o sottoclassi di parole con cui possono combinarsi, come nel caso di *dare* e *darsi*.

Nel primo caso rientrano i verbi che presentano pronomi espletivi o riflessivi desemantizzati, che sono lessicalizzati nel verbo (Schwarze 2009:143), e.g., *arrabbiarsi*, *vergognarsi*, e le cui forme senza clitico non esistono se non nelle costruzioni causative, e.g., *fece arrabbiare Maria*.

I verbi inerentemente clitici, categoria di cui Viviani (2006) presenta un'interessante analisi dal punto di vista sintattico e lessicografico, sono caratterizzati dalla co-occorrenza di un verbo in combinazione con uno o più clitici non riflessivi a rappresentare la pronominalizzazione di uno o più complementi. Come sostenuto da Berruto (1987), alcuni verbi mostrano la tendenza a portare con sé un clitico desemantizzato il cui valore è puramente rafforzativo. A partire dalla definizione di procomplementari di De Mauro (1999) possiamo dire, seguendo Viviani (2006), che in questa categoria rientrano i verbi che accettano uno o due elementi pronominali e che presentano tra gli elementi caratterizzanti la possibilità di produrre poliremathiche, spesso distanti dal valore semantico della base procomplementare, come in *sentire, sentirsi, sentirci, sentirsela.* In questa categoria rientrano anche le costruzioni definite da Masini (2012) verbo-pronominali intensive (CVP) che presentano, quindi, un doppio pronome, come nel caso di *andarsene*, *cavarsela*. In cui, come nota Schwarze (2012), il clitico *si* non realizza nessuna funzione grammaticale nè rappresenta un riferimento anaforico al soggetto ma piuttosto una marca grammaticale (Jezek, 2005:252).

Infine, la categoria dei VID include tutte le espressioni idiomatiche e quasi idiomatiche verbali che presentano clitici. Si possono distinguere due tipologie di comportamenti in queste categorie: l'invariabilità di genere e/o numero (per i clitici che non rappresentano pronomi personali) e l'invariabilità della persona (generalmente alla terza persona singolare o plurale). Come ricordato da Ballarin e Nitti (2019), Serianni (2016:254) indica anche alcuni impieghi che danno vita a costruzioni quasi idiomatiche, in quanto il valore del clitico risulta quasi non riconoscibile, e.g., *averne a male*, *valerne la pena*. Inoltre, in alcune di queste forme sono pre-

senti elementi che confermano la struttura ellittica "in cui si sottintende un sostantivo" (ivi), e.g., *prenderne di santa ragione*, *dirne di tutti i colori*.

4. Formalizzazione dei clitici nelle espressioni poliremiatiche verbali

A partire dagli studi teorici sulle espressioni poliremiatiche, molto è stato fatto nell'ambito della formalizzazione di questi fenomeni linguistici (Copestake *et al.*, 2002; Grégoire, 2007). In particolare, le ricerche si sono concentrate sulla definizione degli aspetti problematici nella rappresentazione di tali espressioni, sia in termini di recupero informazioni sia in termini di riutilizzabilità delle risorse create, al fine di determinare quali informazioni debbano essere registrate e rappresentate nello sviluppo di lessici e risorse terminologiche per scopi di TAL (si veda a tal proposito l'analisi di Escartín *et al.*, 2013).

Nell'ambito delle ricerche più recenti sullo sviluppo di risorse ontologiche per l'italiano, Kahn *et al.* (2014) propongono la pubblicazione del lessico italiano in linked open data, mentre Kahn *et al.* (2018) applicano il *Semantic Web Rule Language* (SWRL)[15] per integrare in un dataset LOD regole di flessione per i verbi monorematici; Racioppa e Declerck (2019) sperimentano l'applicazione di regole morfo-sintattiche alle risorse del Open Multilingual Wordnet, proponendo un sistema di mappatura semantica con gli elementi del modello OntoLex-*lemon*.

In questo lavoro, l'approccio scelto per la formalizzazione si basa sull'utilizzo del modello OntoLex-*lemon*, per il quale l'OntoLex Community Group sta sviluppando un sistema di regole per la generazione/riconoscimento delle parole[16], che consente di definire una funzione che prenda in ingresso una stringa e restituisca un valore booleano sulla base di una corrispondenza con un certo pattern, rappresentando tutti i possibili cammini, cioè i collegamenti tra coppie di vertici, all'interno di un grafo. Il processo di formalizzazione si svolge in due momenti: nel primo si

[15] https://www.w3.org/Submission/SWRL/

[16] Si fa riferimento ai lavori dell'OntoLex Community Group per lo sviluppo del modulo morfologico cui l'autrice contribuisce attivamente.

provvede alla creazione di regole descrittive degli aspetti morfo-sintattici e fonologici/ortografici dei clitici, nel secondo si descrivono e formalizzano i pattern delle espressioni polirematiche verbali che ospitano tali clitici.

4.1. Regole descrittive

Le caratteristiche dei clitici nelle espressioni polirematiche verbali individuate precedentemente vengono formalizzate secondo diversi livelli: un **livello sintattico** che pertiene all'ordine lineare dei costituenti e al fenomeno dei cumuli di clitici, un **livello morfologico** in cui si prendono in considerazione le restrizioni di genere/numero/caso dei clitici in queste costruzioni, un livello **fonologico/ortografico** in cui si descrivono fenomeni come la dissimilazione**.**

Livello sintattico (LS). In questa fase iniziale di sviluppo delle regole e della relativa formalizzazione, a livello sintattico prendiamo in considerazione solo gli aspetti relativi alle posizioni che un clitico, o un gruppo di clitici, può assumere rispetto a un verbo ospite e le regole che governano i gruppi di clitici.

Alcune combinazioni sono ovviamente limitate dalle proprietà sintattiche dei verbi e, quindi, dalla loro struttura argomentale. In tale trattazione non vengono considerati questi aspetti combinatori, tranne nei casi in cui le combinazioni, pur non escluse dalla sintassi del verbo, non sono considerate ammissibili per vincoli relativi alle combinazioni dei cumuli di clitici. Inoltre, la posizione del clitico nel contesto frastico può essere soggetta anche a scelte prosodiche e comunicative che danno vita a fenomeni come l'anteposizione, la dislocazione, la topicalizzazione. Tali fenomeni non sono oggetto di trattazione in questo lavoro in quanto pertengono ad un livello di analisi più ampio sulle strutture frastiche.

Le regole descrittive a questo livello riguardano, quindi, l'ordine lineare dei costituenti e i cumuli di clitici.

A. Ordine lineare dei costituenti. La posizione normale del clitico è preverbale con i verbi di modo finito e postverbale con quelli di modo non finito e l'imperativo positivo. Con

l'imperativo negativo, la scelta tra proclisi o enclisi è opzionale, così come avviene nelle perifrasi *star* + gerundio o infinito, in cui la scelta può sottendere una motivazione prosodica. In presenza di ausiliari del verbo, di *fare* causativo e di *lasciare,* si verifica la proclisi del clitico anche sul secondo verbo che diventa così il verbo ospite, e.g., *mi fece specchiare*.

B. Cumulo di clitici. I clitici in italiano possono disporsi in sequenza, creando appunto quelli che vengono definiti cumuli. L'ordine dei clitici nei cumuli può avere un ordine diverso rispetto a quello dei rispettivi costituenti a cui fanno riferimento nel contesto frastico. Cordin & Calabrese (1998) definiscono l'ordine dei clitici cumulati in base al *rango* cui appartengono secondo uno schema di priorità. Il numero dei clitici di cui può essere formato un cumulo è in genere limitato alla combinazione di due elementi, ma la combinazione di tre elementi non è agrammaticale come in *gli ce ne volle del bello e del buono* (Cordin & Calabrese 1988: 591).

Sulla base di questi elementi è possibile creare una serie di regole descrittive che verranno poi utilizzate per la formalizzazione in Onto-Lex-*lemon*. Così, ad esempio, per descrivere il fenomeno dell'enclisi del clitico con verbi all'imperativo, infinito, gerundio (non perifrastico) e participio si utilizza la regola LS2 (Tabella 1).

Livello morfologico (LM). I tratti di numero, genere e caso codificati nei clitici operano direttamente al livello della frase, del testo e della situazione e, per tale motivo, soprattutto in certe occorrenze idiomatiche o quasi idiomatiche, sono soggetti a restrizioni particolari.

A. Tratti di numero/genere. La possibilità di specificare il paradigma flessivo consente da un lato il riutilizzo di risorse morfologiche esistenti, dall'altro di specificare le restrizioni di numero e genere dell'elemento clitico, che danno vita a diversi fenomeni, come ad esempio il cambio di significato, e.g., *prenderle* e *prenderla*, e il cambio della struttura argomentale del verbo ma solo con quelle restrizioni specifiche, e.g., *spas-*

sarsi (intransitivo), *spassarsela* (transitivo), **spassarselo*. Generalmente, i tratti di numero/genere dei pronomi clitici sono influenzati dal sintagma nominale che è in rapporto anaforico con il pronome; quando il riferimento anaforico non è rintracciabile, come ad esempio nel caso di *spassarsela*, le costruzioni presentano restrizioni specifiche sugli elementi clitici, dando vita a espressioni idiomatiche e quasi idiomatiche.

B. Tratti di caso. Questi tratti, connessi al tipo di funzione grammaticale dell'argomento che il clitico espleta e alla struttura argomentale del verbo, possono essere soggetti a restrizioni combinatorie.

Le regole descrittive per il livello morfologico, presentate nella Tabella 2, vengono quindi utilizzate per formalizzare ad esempio gli accordi di genere e numero col verbo ospite (LM1 in Tabella 2) ed eventuali vincoli morfologici che contribuiscono a definire i fenomeni.

Livello fonologico/ortografico (LP). A questo livello, come evidenziato nella Tabella 3, vengono formalizzate le descrizioni dei mutamenti fonologici e ortografici, come il raddoppiamento della consonante del clitico negli imperativi monosillabici (LP1 in Tabella 3), la cancellazione della *e* finale nell'infinito seguito da un clitico, i diversi tipi di dissimilazione che avvengono nei cumuli di clitici e le restrizioni combinatorie (Cordin & Calabrese, 1988).

4.2. Pattern delle espressioni polirematiche verbali

Le regole descrittive relative agli aspetti morfo-sintattici e fonologici/ortografici dei clitici sono utilizzate per formalizzare i pattern delle espressioni polirematiche verbali che presentano tali particelle.

I pattern delle diverse espressioni verbali sono stati descritti a partire dall'osservazione delle occorrenze estratte dal corpus PARSEME-It Verbal MultiWord Expressions (VMWE)[17] (Monti & di Buono, 2019),

[17] https://github.com/UNIORNLP/PARSEME-It-Corpus

annotato con le diverse classi descritte in precedenza, e, successivamente, integrate con le relative informazioni presenti nel GRADIT (De Mauro, 1999). Sulla base di questi pattern è stato creato un sistema di regole combinatorie, di restrizione e co-occorrenza in grado di rappresentare il complesso sistema dei clitici nelle espressioni polirematiche verbali dell'italiano. Questo sistema di regole viene formalizzato in OntoLex-*lemon,* integrando le regole descrittive dei clitici, per garantire l'efficacia della generazione/riconoscimento delle espressioni verbali sulla base dei loro pattern descrittivi.

La formalizzazione si basa sulla scomposizione dell'entrata lessicale nei suoi costituenti morfologici, non minimi, e con la specifica del set di regole necessarie a creare il paradigma flessivo dell'infinito per quel dato pattern. Questo tipo di formalizzazione consente di riutilizzare le regole descritte per altre entrate che presentano lo stesso pattern.

I pattern identificati sono presentati nella Tabella 4; la formalizzazione di tali pattern viene presentata, per motivi di sintesi, soltanto per l'entrata *prendersela* (Figura 2).

L'entrata di esempio, caratterizzata dalla presenza di un cumulo di due clitici, di cui il primo pronome personale e il secondo oggetto diretto al femminile singolare *si* e *la*, appartiene al gruppo di polirematiche, che presentano restrizioni di numero e genere sugli elementi clitici, descritte col pattern VP4 (Tabella 4). Le restrizioni di numero e genere vincolano soltanto il secondo elemento clitico mentre il primo segue la flessione del verbo ospite. Tale restrizione morfologica deriva dalla possibilità, data la natura transitiva del verbo ospite, di una modifica di significato qualora il secondo clitico presentasse altre marche di numero e genere, come in *me lo prese*, riconducibile a una diversa entrata lessicale, i.e., *prendere*.

Nella forma all'infinito, lemmatizzata come `ontolex: LexicalEntry`, viene formalizzata l'enclisi degli elementi clitici (Regola LS2 - Tabella 1), la caduta della *e* per la presenza di tali elementi (Regola LP2 - Tabella 3) e una dissimilazione della *i* del pronome personale *si* dovuta al suo co-occorrere in un cumulo con un secondo clitico all'accusativo *la* (Regola LP3 - Tabella 3). I comportamenti flessivi dei due clitici sono differenti: il pro-

nome personale *si* segue la flessione del verbo ospite (Regola LM1 - Tabella 2), mentre il secondo clitico, come detto in precedenza, presenta restrizioni di numero singolare e genere femminile (Regola LM3 - Tabella 2).

Per la rappresentazione delle altre forme flesse ai modi finiti, le regole morfologiche e fonologiche/ortografiche per il cumulo di clitici vengono mantenute, mentre variano quelle sintattiche, in quanto gli elementi clitici subiscono un fenomeno di proclisi (Regola LS1 - Tabella 1). A queste, si aggiungono le regole di flessione del verbo senza elementi clitici[18], che si riferiscono alla flessione standard del verbo ospite in *-ere*.

5. Conclusioni

In questo lavoro è stata presentata una proposta di rappresentazione formale dei clitici nelle le espressioni polirematiche verbali dell'italiano, basata su una analisi degli aspetti sintattici, morfologici e fonologici/ortografici che governano il sistema dei clitici pronominali e dei verbi ospite. Tale analisi ha consentito lo sviluppo di un sistema di regole descrittive, formalizzate utilizzando il modello OntoLex-*lemon*, che garantiscono una rappresentazione delle informazioni adeguata alla creazione di risorse linguistiche interoperabili.

La presente ricerca approfondisce, in primo luogo, le attuali possibilità di formalizzazione su base ontologica di risorse linguistiche per l'italiano, che possano supportare lo sviluppo di applicazioni monolingue e multilingue per il TAL ed essere integrate nelle tecnologie semantiche per il linguaggio. Inoltre, i risultati di tale formalizzazione possono contribuire alla ricerca linguistica sul sistema dei clitici in italiano e alla valutazione della capacità rappresentativa del modello, al fine di garantire lo sviluppo di elementi e moduli compatibili con le esigenze di rappresentazione della lingua italiana.

[18] Tali regole non sono presentate in questo lavoro per motivi di sintesi.

Ringraziamenti

Questo lavoro è stato supportato dal Programma Operativo Nazionale Ricerca e Innovazione 2014-2020 - Fondo Sociale Europeo, Azione I.2 "Attrazione e Mobilità Internazionale dei Ricercatori" Avviso D.D. n 407 del 27/02/2018.

Riferimenti bibliografici

Ballarin E., and Nitti P., 2019, "Le funzioni del clitico "ne" nelle interlingue postbasiche di apprendenti l'italiano L2, a un anno di distanza dall'immersione nel contesto italofono", *Italiano Lingua Due*, 11.2: 62-76.

Benincà P.; Salvi G. and Frison L., 1988, "L'ordine degli elementi della frase e le costruzioni marcate", *Grande grammatica italiana di consultazione*, 1: 115-225.

Bergenholtz H., and Tarp S., 2005, "Dictionaries and inflectional morphology", *Encyclopedia of Language and Linguistics*, Oxford, Pergamon Press, 577-580.

Berretta M., 1984, *Per uno studio dell'apprendimento dell'italiano in contesto naturale: il caso dei pronomi personali atoni*, Bologna, il Mulino.

Berruto G., 1987, *Sociolinguistica dell'italiano contemporaneo*, Vol. 33, Roma, Carocci.

Bird S. and Liberman M., 1998, "Towards a formal framework for linguistic annotations", in Mannell R.H. and Jordi Robert-Ribes J. (eds.), *Fifth International Conference on Spoken Language Processing, CSLP-1998*, paper 0774.

Budanitsky A., and Hirst G., 2006, "Evaluating wordnet-based measures of lexical semantic relatedness", *Computational Linguistics*, 32.1: 13-47.

Buitelaar P., *et al*., 2006, "Linginfo: Design and applications of a model for the integration of linguistic information in ontologies", in *Proceedings of the OntoLex Workshop at LREC 2006*, European Language Resources Association (ELRA), 28-34.

Cennamo M., 1993, *The reanalysis of reflexives: a diachronic perspective*, Napoli, Liguori.

Chiarcos C., *et al.*, 2013, "Towards open data for linguistics: Linguistic linked data", *New Trends of Research in Ontologies and Lexical Resources*, Berlin, Heidelberg, Springer, 7-25.

Chomsky N., 1965, *Aspects of the theory of syntax*, Cambridge, Multilingual Matters, MIT Press.

Cimiano P.; P. McCrae J. and Buitelaar P., 2016, *Lexicon Model for Ontologies: Community Report*, Community Group Final Report, W3C.

Cimiano P., *et al.*, 2007, "LexOnto: A model for ontology lexicons for ontology-based NLP", in Aberer K. *et al.* (eds.), *Proceedings of the OntoLex 07 Workshop held in conjunction with ISWC'07*, 81-92.

Cinque G., 1988, "On si constructions and the theory of arb", *Linguistic inquiry* 19.4 : 521-581.

Copestake A.; Lambeau F.; Villavicencio A.; Sag I.; Bond F.; Baldwin T. and Flickinger D., 2002, "Multiword expressions: Linguistic precision and reusability", in Rodríguez M.G. and Araujo C.P.S. (eds.), *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, European Language Resources Association (ELRA), 1941-1947.

Cordin P., 2001, "I pronomi riflessivi", in Reni L. , Salvi G. & Cardinaletti A. (a cura di), *Grande grammatica italiana di consultazione*, Bologna, il Mulino, 1: 607-617.

Cordin P. & Calabrese A., 1988, "I pronomi personali", in Renzi L., Salvi G. & Cardinaletti A. (a cura di), *Grande grammatica italiana di consultazione*, Bologna, il Mulino, 1: 535-592.

De Alencar L.F.; Kelling C., 2005, "Are reflexive constructions transitive or intransitive? Evidence from German and Romance", in Butt M. and King T.H. (eds.), *Proceedings of the LFG05 Conference*, Stanford, CSLI Publications, 1-20.

De Mauro T., 1999-2000, *Grande Dizionario Italiano dell'Uso*, Torino: UTET, 6 vol. (e CD-ROM); Id., *Nuove parole italiane dell'uso*, ibid. 2003; Id., *Nuove parole italiane dell'uso II*, ibid. 2007 (e Supporto Digitale – Penna USB).

Eckart R., 2008, "Choosing an XML database for linguistically annotated corpora", *Sprache und Datenverarbeitung* 32.1: 7-22.

Escartín C.P., *et al.*, 2013, "Representing multiword expressions in lexical and terminological resources: an analysis for natural language processing purposes", in Kosem I. *et al. (eds.), Proceedings of eLex 2013*, Brno, Lexical Computing, 338-357.

Grégoire N., 2007, "Design and implementation of a lexicon of Dutch multiword expressions", in Gregoire N. *et al.* (eds.), *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions,* Stroudsburg, Association for Computational Linguistics, 17-24.

Ježek E., 2005, *Lessico: classi di parole, strutture, combinazioni*, Bologna, il Mulino.

Khan F., and Frontini F., 2014, "Publishing PAROLE SIMPLE CLIPS as Linguistic Linked Open Data", in Basili R., Lenci A., Magnini B. (eds.), *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014*, Pisa, Pisa University Press, 224-228.

Khan F., *et al.*, 2018, "SWRL your lexicon: adding inflectional rules to a LOD", in *Abstract of The XVIII EURALEX International Congress,* 68.

Klimek B., *et al.*, 2019, "Challenges for the Representation of Morphology in Ontology Lexicons", in Kosem I. *et al.* (eds.), *Electronic lexicography in the 21st century, Proceedings of the eLex 2019 conference,* Brno, Lexical Computing, 570-591.

Lo Cascio V., 1970, *Strutture pronominali e verbali italiane*, Bologna, Zanichelli.

Masini F., 2012, "Costruzioni verbo-pronominali "intensive" in italiano", Language and the brain–Semantics., in Bambini V.; Ricci I.; Bertinetto P.M. (eds.), *Linguaggio e cervello - Semantica / Language and the brain - Semantics - Atti del XLII Congresso Internazionale di Studi della Società di Linguistica Italiana (Pisa, SNS, 2008)* (SLI), Roma, Bulzoni, 1 - 22.

McCarthy D., and Navigli R., 2009, "The English lexical substitution task", *Language Resources and Evaluation,* 43.2: 139-159.

McCrae J.P., *et al.*, 2012, "Interchanging lexical resources on the Semantic Web", *Language Resources and Evaluation,* 46.6:701–709

McCrae J.P., *et al.*, 2017, "The Ontolex-Lemon model: development and applications", in Kosem I. *et al.* (eds.), *Proceedings of eLex 2017 conference,* Brno, Lexical Computing, 587-597.

McGuinness D.L. and van Harmelen F., 2004, *OWL Web Ontology Language Overview* (technical report), W3C Recommendation .

Monti J. and di Buono M.P., 2019, "PARSEME-It: an Italian corpus annotated with verbal multiword expressions", *IJCoL - Italian Journal of Computational Linguistics,* 5:61-94.

Montiel-Ponsoda E., *et al.*, 2011, "Enriching ontologies with multilingual information", *Natural language engineering,* 17.3: 283-309.

Navigli R., 2009, "Word sense disambiguation: A survey", *ACM computing surveys (CSUR),* 41.2: 1-69.

Navigli R., and Ponzetto S.P., 2012, "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network", *Artificial Intelligence,* 193: 217-250.

Navigli R., 2006, "Meaningful clustering of senses helps boost word sense disambiguation performance", in Calzolari N. *et al.* (eds.), *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, Stroudsburg, Association for Computational Linguistics, 105-112.

Neely J.H.; Keefe D.E. and Ross K.L., 1989, "Semantic priming in the lexical decision task: Roles of prospective prime-generated expectancies and retrospective semantic matching", *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 15.6: 1003.

Niemann E. and Gurevych I., 2011, "The people's web meets linguistic knowledge: automatic sense alignment of Wikipedia and Wordnet", in Bos J. and Pulman S. (eds.), *Proceedings of the Ninth International Conference on Computational Semantics*, Stroudsburg, Association for Computational Linguistics, 205-214.

Odijk J., 2004, "A proposed standard for the lexical representation of idioms", in Williams G. and Vessier S. (eds.), *EURALEX 2004 proceedings*, Lorient, Université Bretagne Sud, 153-164.

Penello N., 2004, "I clitici locativo e partitivo nelle varietà italiane settentrionali", *Quaderni di lavoro dell'ASIS,* 4: 37-103.

Perlmutter D.M., 1978, "Impersonal passives and the unaccusative hypothesis", *Annual meeting of the Berkeley Linguistics Society*, 4:157-189.

Pilehvar M.T.; Jurgens D. and Navigli R., 2013, "Align, disambiguate and walk: A unified approach for measuring semantic similarity", in Schuetze H.; Fung P.; Poesio M. (eds.), *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* Stroudsburg, Association for Computational Linguistics, 1341-1351.

Pilehvar M.T. and Navigli R., 2014, "A robust approach to aligning heterogeneous lexical resources", in Toutanova K.; Wu H. (eds.), *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Stroudsburg, Association for Computational Linguistics, 468-478.

Racioppa S. and Declerck T., 2019, "Enriching Open Multilingual Wordnets with Morphological Features", in Bernardi R.; Navigli R.; Semeraro G. (eds.), *Proceedings of the Sixth Italian Conference on Computational Linguistics. Italian Conference on Computational Linguistics (CLIC-it-2019)*, Bari, CEUR 10/2019, 228–233.

Salvi G. and Vanelli L., 2004, *Nuova grammatica italiana*, Bologna, il Mulino.

Savary A., *et al.*, 2015, "PARSEME–PARSing and Multiword Expressions within a European multilingual network", in *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*.

Schierholz S.J., 2015, "Methods in lexicography and dictionary research", *Lexikos,* 25: 323-352.

Swanepoel P.H., 2015, "The design of morphological/linguistic data in L1 and L2 monolingual, explanatory dictionaries: a functional and/or linguistic approach?", *Lexikos* 25: 353-386.

Schwarze C., 2009, *Grammatica della lingua italiana*, a cura di Colombo A., Roma, Carocci.

Schwarze C., 2012, *Romance clitic pronouns in lexical paradigms*, Amsterdam, John Benjamins.

Serianni L., 1991, *Grammatica italiana. Italiano comune e lingua letteraria*, Torino, UTET.

Simone, Raffaele, 1983, "Punti di attacco dei clitici in italiano", in Albano Leoni F.; Gambarare D.; Lo Piparo F.; Simone R. (eds.), *Italia linguistica: idee, storia, strutture*, Bologna, il Mulino, 285-307.

Snow R., *et al.*, 2007, "Learning to merge word senses", in Eisner J. (ed.), *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (emnlp-conll)* Stroudsburg, Association for Computational Linguistics, 1005-1014.

Turney P.D. and Pantel P., 2010, "From frequency to meaning: Vector space models of semantics", *Journal of artificial intelligence research,* 37: 141-188.

Viviani A., 2006, *I verbi procomplementari tra grammatica e lessicografia*, Firenze, Le Lettere.

Appendix

Nome	**Descrizione**	**Esempio**
LS1	Proclisi del clitico con verbi all'indicativo o congiuntivo[19]	*Si* specchia
LS2	Enclisi del clitico con verbi all'imperativo, infinito, gerundio (non perifrastico), participio	Specchiar*si*
LS3	Proclisi del clitico in presenza di ausiliari del verbo, *fare* causativo, *lasciare*	*Si* è specchiato
LS4	Proclisi del clitico in presenza di *fare* causativo e *lasciare* con ausiliari come *avere*	*Mi* fece specchiare

[19] Questa regola include anche l'imperativo alla terza persona, incluso la forma di cortesia, in cui il verbo è al congiuntivo.

LS5	Proclisi/enclisi (opzionale) con verbi all'imperativo negativo	Non specchiar*ti* Non *ti* specchiare
LS6	Proclisi/enclisi del clitico nelle perifrasi *star* + gerundio o infinito	Sta specchiando*si* *Si* sta specchiando

Tabella 1. *Regole per il livello sintattico*

Nome	**Descrizione**	**Esempio**
LM1	Accordo di genere e numero col verbo ospite	*Mi* specchio
LM2	Vincolo di genere e numero del clitico con funzione accusativa femminile plurale	*Se le* danno
LM3	Vincolo di genere e numero del clitico con funzione accusativa femminile singolare[20]	Se *la* spassano
LM4	Forme plurali obbligatorie del clitico	prender*le*
LM5	Forme singolari obbligatorie	spassar*sela*
LM6	Cambio di significato in base al numero del clitico	Dar*le* vs. dar*la*
LM7	Cambio di significato per la presenza del clitico	Fars*i* vs. fare
LM8	Cumulo di clitici, cambio di significato e di struttura argomentale (da intransitivo a transitivo) per la presenza del secondo clitico oggetto diretto	Spassars*i* vs. spassar*sela*

Tabella 2. *Regole per il livello morfologico*[21]

[20] Questa regola si applica anche a verbi la cui struttura tematica si modifica in base alla presenza del secondo clitico e di cui esistono due forme lemmatizzate (intransitivo *spassarsi* vs. transitivo *spassar*sel*a*)

[21] Si noti che per le regole LM6, LM7 e LM8 non vengono formalizzati i diversi significati assunti dalle singole entrate, bensì la presenza del fenomeno di cambio del si-

Nome	Descrizione	Esempio
LP1	Raddoppiamento della consonante iniziale con imperativi monosillabici	Fa*tti*
LP2	Caduta della *e* finale nelle forme infinite	Far*si*[22]
LP3	Dissimilazione della *i* davanti ai clitici accusativi e genitivo	Prender*sela*
LP4	Dissimilazione nel cumulo si si	*Ci si* conosce
LP5	Dissimilazione nel cumulo vi vi	*Vi ci* incontrerete
LP6	Dissimilazione nel cumulo ci ci	Ci vedremo lì
LP7	Dissimilazione e ordine dei clitici nei cumuli a tre elementi	*Gli ce ne* volle del bello e del buono

Tabella 3. *Regole per il livello fonologico/ortografico*

Nome	Descrizione	Esempio
VP1	Espressioni con clitico con flessione in accordo col verbo ospite	Specchiar*si*
VP2	Espressioni con clitico con flessione in accordo col verbo ospite + preposizione e avverbio fissi	Prender*si* a male
VP3	Espressioni con cumulo di clitici di cui il primo con flessione in accordo col verbo ospite e il secondo con restrizioni di genere e numero + preposizione e avverbio fissi	Prender*sela* a male
VP4	Espressioni con cumulo di clitici di cui il primo con flessione in accordo col verbo e dissimilazione, il secondo oggetto con restrizioni di genere e numero	Dar*sele*

gnificato, causato dalle caratteristiche morfo-sintattiche dell'elemento clitico coinvolto nell'espressione polirematica.

[22] Questa regola si applica ai verbi lemmatizzati sia nella forma con il clitico che nella forma senza clitico, come per le occorrenze che mostrano un cambio di significato (*fare* vs. f*arsi*)

VP5	Espressioni con verbi fattivi o percettivi in costruzioni causative con risalita del clitico con funzione locativa	Arrivar*ci* In *Ci* fece arrivare
VP6	Espressione con cumulo di clitici di cui il primo con funzione locativa e il secondo con flessione in accordo col verbo	Metter*cisi*
VP7	Espressione con clitico partitivo + preposizione, aggettivo, determinante e nome non modificabili e fissi	Far*ne* di tutti i colori
VP8	Espressioni con cumulo di clitici di cui il primo con flessione in accordo col verbo e dissimilazione, il secondo oggetto con restrizioni di genere e numero	Meritar*sela* vs. meritar*sele*
VP9	Espressione verbale con clitico locativo + avverbio	Veder*ci* chiaro
VP10	Espressione verbale con clitico con flessione in accordo col verbo + avverbio fisso	Sentirs*i* male
VP11	Espressione verbale con cumulo di clitici senza flessione, di cui il primo con funzione dativa, il secondo partitivo	Cantar*gliene* Dar*gliene*

***Tabella 4**. Descrizioni dei pattern delle VMWE*

```
#lemmatizzazione dell'entrata lessicale pren-
dersela
:lex_prendersela a ontolex:LexicalEntry ;
    lexinfo:partOfSpeech lexinfo:mainVerb ;
    ontolex:canonicalForm [ ontolex:writtenRep
«prendersela»@it ] ;
    ontolex:morphologicalPattern :it-VP4 .

#scomposizione dell'entrata nei suoi costituen-
ti morfologici
:form_prendersi a ontolex:Form ;
morph:consistOf :it-VP4_const
morph:rootMorph  [ ontolex:writtenRep "prend"@
it ]
```

```
morph:affixMorph [ ontolex:writtenRep “ere”@it ]
:morph_si a morph:partMorph [ ontolex:written-
Rep “si”@it ]
:morph_la a morph:partMorph [ ontolex:written-
Rep “la”@it ]

#definizione del paradigma flessivo
:it-VP4 a morph:Paradigm ;
rdfs:comment “Espressione con una testa verbale
seguita da un cumulo di clitici di cui il primo
con flessione e dissimilazione e il secondo oggetto
diretto non modificabile nel numero e nel genere” .

#elenco delle regole LS, LM, LP per l’entrata VP4
:it-VP4_2_type_infin a morph:SubParadigm ;
morph:paradigm :it-VP4 .
:it-CLI_LS2 a morph:SubParadigm ;
morph:paradigm :it-cli .
:it-CLI_LM1 a morph:SubParadigm ;
morph:paradigm :it-cli
:it-CLI_LM3 a morph:SubParadigm ;
morph:paradigm :it-cli
:it-CLI_LP2 a morph:SubParadigm ;
morph:paradigm :it-cli
:it-CLI_LP3 a morph:SubParadigm ;
morph:paradigm :it-cli

#realizzazione delle regole per l’infinito -
output: prendersela
:it-VP4_2_type_infin a morph:Rule ;
morph:subParadigmOf :it-VP4 ;
morph:inflectsFor [lexinfo:tense lexinfo:present ;
lexinfo:mood lexinfo:infinitive];
rdfs:label “”@it ;
morph:replacement [morph:source “$”; morph:tar-
get “”] .

#regola LS2 - enclisi del clitico
:it-CLI_LS2 a morph:Rule ;
morph:subParadigmOf :it-cli ;
```

```
morph:inflectsFor [lexinfo:tense lexinfo:present
;
lexinfo:mood lexinfo:infinitive];
rdfs:label “”@it ;
morph:replacement [morph:source “^$”; mor-
ph:target “si”] .

#regola LM1 - accordo genere e numero col verbo
ospite per il primo clitico
:it-CLI_LM1 a morph:Rule ;
morph:subParadigmOf :it-cli ;
morph:inflectsFor [lexinfo:number lexinfo:singu-
lar ;
lexinfo:person lexinfo:third];
rdfs:label “”@it ;
morph:replacement [morph:source “$”; morph:tar-
get “”] .

#regola LM3 - restrizioni di genere e numero
per il secondo clitico
:it-CLI_LM3 a morph:Rule ;
morph:subParadigmOf :it-cli ;
morph:inflectsFor [lexinfo:number lexinfo:singu-
lar ;
lexinfo:gender lexinfo:feminine];
rdfs:label “”@it ;
morph:replacement [morph:source “$”; morph:tar-
get “la”] .

#regola LP2 - caduta della e finale della forma
all’infinito per la presenza di un clitico in
posizione enclitica
:it-CLI_LP2 a morph:Rule ;
morph:subParadigmOf :it-cli ;
morph:inflectsFor [lexinfo:tense lexinfo:present
;
lexinfo:mood lexinfo:infinitive];
rdfs:label “”@it ;
morph:replacement [morph:source “^(.)ere$”;
morph:target “er”] .
```

```
#regola LP3 - dissimilazione della i del primo
clitico per la presenza nel cumulo di un cliti-
co con funzione accusativa
:it-CLI_LP3 a morph:Rule ;
morph:subParadigmOf :it-cli ;
morph:inflectsFor [lexinfo:number lexinfo:singu-
lar ;
lexinfo:gender lexinfo:feminine];
rdfs:label “”@it ;
morph:replacement [morph:source “si”; mor-
ph:target “se”] .
```

***Figura 2**. Esempio di formalizzazione dell'entrata lessicale prendersela*