

TOPOI

4

*Direttori*

Fernando MARTÍNEZ DE CARNERO CALZADA

Sapienza Università di Roma

Luisa Allesita MESSINA FAJARDO

Università degli Studi Roma Tre

*Comitato scientifico*

Juan Carlos ABRIL

Universidad de Granada

Maria Pilar Agustina CAPANAGA CABALLERO

Alma Mater Studiorum – Università di Bologna

Marina FERNÁNDEZ LAGUNILLA

Universidad Complutense de Madrid

Trinis Antonietta MESSINA FAJARDO

Università degli Studi di Enna “Kore”

Oana SALISTEANU

Universitatea din București

Antonio RICO SULAYES

Universidad de las Américas Puebla

*Comitato redazionale*

Mariarosaria COLUCCIELLO

Università degli Studi di Salerno

Cosimo DE GIOVANNI

Università degli Studi di Cagliari

Paolo RONDINELLI

Accademia della Crusca

Alessia Anna Serena RUGGERI

Università degli Studi Roma Tre

TOPOI



La collana accoglie studi, testi e raccolte di saggi dedicati all'analisi dei luoghi comuni da un punto di vista interdisciplinare e interculturale, spaziando dalla linguistica alla letteratura, dai linguaggi settoriali alle forme dello stile. La topica si rivela, all'interno della tradizione culturale, filosofica e letteraria, come uno strumento essenziale per la trasmissione del pensiero. Il suo contributo alla costruzione del senso si manifesta attraverso un ampio repertorio di generi discorsivi, come i proverbi, gli aforismi, gli emblemi e coinvolge anche molte aree del sapere: diritto, religione, politica, medicina, economia. L'utilità di questo tipo di approccio, ampio e globale, verso lo studio dei luoghi comuni, peraltro di grande importanza per una più approfondita comprensione dei diversi periodi storici, consiste innanzitutto nell'offrire uno strumento d'indagine con il quale la ricerca si apre a nuove prospettive.

*Web content*



Volume published with the contribution and the cooperation of Associazione Italiana di Fraseologia e Paremiologia Phrasis.

Antonio Rico–Sulayes

**Authorship Attribution  
on Crime–Related Social Media**

Research on the Darknet in Forensic Linguistics

*Prefazione di*  
Luisa A. Messina Fajardo

*Foreword by*  
Ignacio Rodríguez Sánchez





Aracne editrice

[www.aracneeditrice.it](http://www.aracneeditrice.it)  
[info@aracneeditrice.it](mailto:info@aracneeditrice.it)

Copyright © MMXVIII  
Giacchino Onorati editore S.r.l. – unipersonale

[www.giacchinoonoratieditore.it](http://www.giacchinoonoratieditore.it)  
[info@giacchinoonoratieditore.it](mailto:info@giacchinoonoratieditore.it)

via Vittorio Veneto, 20  
00020 Canterano (Rome)  
(06) 45551463

ISBN 978-88-255-1377-6

*No part of this book may be reproduced  
by print, photoprint, microfilm, microfiche, or any other means,  
without publisher's authorization.*

I<sup>st</sup> edition: April 2018

*To the more than 175,000 people  
dead in the fight against organized crime  
Mexico, 2017*



# Table of Contents

- 11 *List of Figures*
- 13 *List of Tables*
- 15 *Prefazione*  
di Luisa A. Messina Fajardo
- 21 *Foreword*  
by Ignacio Rodríguez Sánchez
- 25 **Chapter I**  
*Identifying the Author of Anonymous Documents in Legal Contexts*  
1.1. Why a Book in Authorship Attribution?, 28 – 1.2. Outline for a Research-Oriented Presentation, 36
- 39 **Chapter II**  
*Forensic Linguistics and Authorship Analysis*  
2.1. Tasks of Authorship Analysis, 42 – 2.2. What exactly is Authorship Attribution?, 58 – 2.3. Questions Author Profiling Tries to Answer, 84
- 95 **Chapter III**  
*Authorship Analysis Resources*  
3.1. Between What Features Represent Authorship and How Many, 97 – 3.1.1. *Selection of Features in Authorship Attribution*, 103 – 3.1.2. *Why Reducing Lists of Features and How to Do It?*, 109 – 3.1.3. *Selection of Features in Author Profiling*, 125 – 3.1.4. *Reducing Lists of Features in Author Profiling*, 130 – 3.2. Quantitative Classification in Authorship Analysis, 132 – 3.2.1. *Classifiers for Authorship Attribution*, 138 – 3.2.2. *Classifiers for Author Profiling*, 141 – 3.3. Authorship Analysis Applied on Spanish Data, 143

155	<b>Chapter IV</b> <i>Authorship Attribution of Online Forum Posts in Spanish</i> 4.1. Data for Authorship Attribution Experiments, 157 – 4.1.1. <i>The Drug-Dealing Related Social Media Phenomenon in Mexico</i> , 159 – 4.1.2. <i>Experimental Corpora</i> , 161 – 4.2. Attribution Approach Components, 166 – 4.2.1. <i>Set of Authorship Features</i> , 167 – 4.2.2. <i>Feature Selection</i> , 171 – 4.2.3. <i>Feature Reduction Techniques</i> , 183 – 4.2.4. <i>Tested Classifiers</i> , 193 – 4.2.5. <i>Attribution Approach Configurations</i> , 208
213	<b>Chapter V</b> <i>Experiments Results</i> 5.1. Measuring Results, 213 – 5.1.1. <i>Success Rate in Experiments: Accuracy</i> , 214 – 5.1.2. <i>Effects of Classifier, Feature Reduction, and Normalization</i> , 217 – 5.1.3. <i>Comparison of Approach Configurations</i> , 221 – 5.1.4. <i>Effects of Automated Tagging and Machine Learning Classification</i> , 226 – 5.1.5. <i>Summary of Results</i> , 228 – 5.2. Linguistic Components in Experiments, 229
247	<b>Conclusion</b> Revisiting the Research Program, 248 – Discussion of Contributions to Authorship Analysis, 250 – What to Do Next?, 254
257	<b>References</b>
273	<b>Glossary</b>

## List of Figures

- 203 Figure 4.1. Linear classification problem with two classes
- 204 Figure 4.2. Non-linear classification problem with two classes
- 206 Figure 4.3. SVMs and the maximum-margin decision
- 221 Figure 5.1. MNB and non-normalized data
- 222 Figure 5.2. NB and non-normalized data
- 224 Figure 5.3. DA and non-normalized data
- 225 Figure 5.4. Results for frequency reduced, non-normalized data



## List of Tables

- 57 Table 2.1. Authorship analysis tasks and related terminology
- 69 Table 2.2. Genre/topic selection in authorship attribution
- 75 Table 2.3. Number of authors in authorship attribution
- 78 Table 2.4. Data units in authorship attribution
- 86 Table 2.5. Categories in author profiling
- 89 Table 2.6. Genre selection in author profiling
- 92 Table 2.7. Experimental data design in author profiling
- 105 Table 3.1. Feature sets in authorship attribution
- 108 Table 3.2. Feature set testing in authorship attribution
- 122 Table 3.3. Feature reduction techniques in authorship attribution
- 127 Table 3.4. Feature arrangement in author profiling
- 129 Table 3.5. Feature set testing by category in author profiling

- 131 Table 3.6. Feature reduction techniques in author profiling
- 139 Table 3.7. Classification methods in authorship attribution
- 142 Table 3.8. Classification methods in author profiling
- 145 Table 3.9. Languages targeted in authorship attribution
- 180 Table 4.1. Frequent Keyboard-Based Emoticons in Spanish
- 182 Table 4.2. Selected authorship features before reduction
- 209 Table 4.3. Approach configurations
- 216 Table 5.1. Confusion matrix for ten authors
- 218 Table 5.2. Results for configurations with MNB
- 219 Table 5.3. Accuracy results for all configurations
- 227 Table 5.4. Best results for 10 authors
- 228 Table 5.5. Approach configurations matching or improving DA
- 232 Table 5.6. Size of feature sets
- 233 Table 5.7. Features frequently selected by IG and CFS
- 245 Table 5.8 Linguistic/paralinguistic features selected by IG or CFS

## Prefazione

di Luisa A. Messina Fajardo<sup>1</sup>

Sono lieta di presentare questo volume che propone, attraverso un approccio originale, uno studio nel campo della fraseologia computazionale. Si tratta, senza dubbio, di un argomento di grande interesse e attualità. Tuttavia, ritenendoci poco edotti in materia, più che addentrarci nella descrizione dettagliata degli argomenti trattati da Antonio Rico-Sulayes, sempre più importanti nella nostra era globale caratterizzata dall'ausilio delle tecnologie informatiche, vogliamo affrontare un argomento a noi più familiare e cioè quello riguardante lo sviluppo degli studi fraseologici. Vogliamo descriverne, anche se brevemente, le tappe più significative a partire dalla nascita dei primi approcci agli studi fraseologici fino a fornire brevi cenni relativi alla fraseologia computazionale.

La fraseologia, considerata sin dai tempi antichi una parente povera della linguistica, dalla fine del XX secolo si è affermata sempre più come una delle tematiche più diffuse e studiate nelle diverse lingue. Le unità fraseologiche (UF) sono oggetto di studio di molti linguisti e di gruppi di ricerca come FRAMESPA (Universidad de Toulouse-Le Mirail, Francia), PHRASEONET (Universidad de Santiago de Compostela), FRASESPAL (Universidad de Santiago de Compostela), FRASEMIA, Fraseología, Paremiología y Traducción (Universidad de Murcia), FRASYTRAM, Fraseología y Traducción Multilingüe (Universidad de Alicante), ALIENTO (INALCO, París; Universidad de Nancy, Francia). A questo proposito occorre ricordare il lavoro

---

<sup>1</sup> Dott.ssa Luisa A. Messina Fajardo è professoressa associata di Lingua, Cultura e Istituzioni dei Paesi di Lingua Spagnola presso l'Università degli Studi Roma Tre. È presidente dell'Associazione Italiana di Fraseologia e Paremiologia.

svolto da alcune associazioni quali PHRASIS (Italia) e EURO-PHRAS (Europa) che attraverso le giornate di studio organizzate e le riviste *Phrasis* e *Paremia* contribuiscono a diffondere e a favorire lo sviluppo degli studi fraseologici a livello internazionale. E ancora, è importantissimo il ruolo svolto in Italia dal Centro Interuniversitario di Geoparemiologia che coordina il progetto di ricerca: Atlante paremiologico Italiano (API).

Negli ultimi trent'anni la fraseologia, anche grazie a queste realtà, è cresciuta: in Europa, oggi, non è più considerata una sotto-disciplina della lessicologia bensì una branca di studi dotata di una sua autonomia. Per quanto riguarda gli approcci allo studio della fraseologia, se fino a pochi anni fa ci si limitava allo studio degli aspetti semantici e testuali delle combinazioni di parole, oggi il campo d'indagine è molto più ampio. Di seguito, presentiamo sette approcci relativi allo studio della disciplina:

*Delimitazione, classificazione, caratterizzazione.* Fanno parte di questa categoria gli studi che hanno come fine ultimo quello di stabilire i limiti della disciplina. Si tratta di un approccio tradizionale in cui gli studiosi cercano di capire quali siano le combinazioni di parole che rientrano nel dominio della fraseologia e quali no. A questo proposito vi sono due correnti di studio contrastanti, a seconda che si consideri la fraseologia attraverso una concezione ampia o ristretta (Corpas Pastor: 1995, 2001); (Ruiz Gurrillo: 1997, 2001), (Zuluaga: 1998); (García-Page Sánchez: 2008). Inoltre, è necessario classificare, caratterizzare e sistematizzare i diversi tipi e sottotipi di unità fraseologiche attraverso l'uso dei criteri di tipo semantico, sintattico, pragmatico e denominativo (Corpas Pastor: 1996, 1998); (Castillo Carballo: 1998); (Zuluaga: 1998); (Ruiz Gurrillo: 1997, 1998, 2001), (Sevilla & Crida: 2015).

Una volta classificate le UF occorre studiare le relazioni che le combinazioni di parole stabiliscono tra loro e con il resto delle unità del sistema della lingua. Si prosegue, pertanto, all'analisi delle restrizioni sintattiche e grammaticali delle combinazioni di parole rispetto ad altre unità del sistema linguistico,

come i composti e le combinazioni libere di parole, i composti e le locuzioni.

*Aspetti pragmatico-testuali.* In questa sezione confluiscono gli studi riguardanti l'uso reale delle combinazioni di parole e le variazioni fraseologiche rispetto alle varietà cronologiche, diafasiche, diatopiche della lingua data. A questo riguardo occorre sottolineare che la creazione della linguistica dei *corpora* ha messo in crisi il concetto di fissazione (*fijación*) (García-Page Sánchez: 1998). Ultimamente, aiutati dalla linguistica computazionale, si sta prestando particolare attenzione alle varianti fraseologiche dialettali e diatopiche.

Importanti, oltre che interessanti, sono gli studi riguardanti le variazioni fraseologiche nella forma scritta e orale di una lingua. A tal proposito è importante evidenziare che quando una combinazione di parole è utilizzata in un testo scritto vengono rispettate le forme canoniche; al contrario, quando invece questa si utilizza in un discorso, subisce delle manipolazioni, modifiche e riduzioni che ne determinano la forma tipica dell'oralità. Gli studi che se ne occupano sono tanti: Corpas Pastor (1996, 1998); García-Page Sánchez (1993); Güell (1999); Vigara Tautste (1998); Nuccorini (2001). Vanno ricordati anche studi che analizzano il significato negativo delle UF scaturito da valori convenzionali (misogini, razziali, sociali) (Calero: 1991, 1998); (Zuluaga: 2001); (Messina Fajardo L.: 2011).

Un altro blocco tematico è costituito dagli studi sulla fraseologia dei linguaggi specialistici, nei quali i termini, le locuzioni terminologiche e le collocazioni costituiscono gli elementi identificativi più caratteristici del discorso specialistico in tutti i suoi livelli (Castillo Carballo: 1999); (Capra: 2008); (Messina Fajardo L. 2012, 2016); (Navarro: 2001).

*Aspetti semantico-semiotici.* La variante semantico-semiotica è tra tematiche più studiate, attualmente, in Europa - insieme a quella terminologica specializzata (giuridica) -; il suo campo d'indagine è quello delle relazioni paradigmatiche (onomasiologie) che le combinazioni di parole stabiliscono tra loro, e in

misura minore, con le altre unità lessicali del sistema linguistico studiato.

Quest'aspetto è molto importante, in quanto, gli studi di combinazioni di parole raggruppate per campi lessico-fraseologici risultano utili per la fraseologia generale e quella comparata.

Sono, altresì, molto importanti gli studi relativi alla polisemia (Mellado: 1998), all'antonimia e alla sinonimia (García-Page Sánchez: 1998); (De Giovanni: 2012), all'iponimia, alla descrizione e comparazione di gruppi tematici (somatismi, Mellado: 1997); (numeri, García-Page Sánchez: 2000); (abbigliamento, Messina Fajardo: 2015) o a campi lessico-fraseologici (Penadés: 2000) e alla ricerca cosciente di sfumature simboliche, etnolinguistiche o culturali.

*Aspetti semantico-cognitivi.* La semantica cognitiva è stata inserita all'interno degli studi fraseologici per stabilire le rappresentazioni mentali soggiacenti le combinazioni di parole (Cacciari: 1986, 1989, 1999, 2001; Cantarini: 1997; Salvador: 1994; Iñesta y Pamies: 2002; Prandi: 2008, 2010, 2013); Messina Fajardo: 2010, 2013). A questo proposito la metafora, come canale di espressione delle emozioni, rappresenta un modello cognitivo di serie fraseologiche e campi fraseologici completi. I significati letterali e idiomatici interagiscono nel lessico mentale, in modo da stabilire una stretta relazione tra la base di motivazione metaforica e il significato unitario delle combinazioni di parole. Le motivazioni metaforiche e simboliche permettono di raggruppare le combinazioni di parole appartenenti a un campo secondo le loro sfere concettuali figurative.

*Aspetti psicolinguistici.* Si cerca di comprendere la realtà psicologica delle unità fraseologiche, quindi, come vengono memorizzate, come avviene la loro elaborazione, la funzione che queste rivestono nell'interazione, ecc. (Lorenzo: 1980); (Corpas Pastor: 2001). Nonostante questi aspetti siano centrali negli studi fraseologici, attualmente, non sono molto sviluppati all'interno della Penisola Iberica.

*Studi di fraseologia comparata.* La fraseologia comparata analizza i sistemi fraseologici di due o più lingue. Quest'analisi racchiude tutti gli aspetti precedentemente menzionati, ma ne fa sorgere di nuovi come gli universali fraseologici, i prestiti linguistici e le corrispondenze interlinguistiche, ricercando le coincidenze di forma e contenuto di combinazioni di parole in lingue diverse. Esempi di queste unità sono gli *europaismi* che si dividono in *europaismi naturali* quando nascono dall'osservazione del mondo che ci circonda, ed *europaismi culturali* quando nascono da fonti comuni della cultura europea. Queste equivalenze vengono classificate su una scala che va dall'equivalenza nulla a quella totale, passando per diverse modalità di equivalenze parziali (Corpas Pastor: 1995, 2000); (Capra: 2012); (Morvay: 1996); (Zuluaga: 1999, 2000, 2001); (Martínez Marín: 1998); (Ruiz Gurrillo: 1994, 1995); (Navarro: 2008) ecc.

*Studi di fraseologia in rapporto con le Nuove Tecnologie e le TICs.* Sono molte le possibilità che offrono gli studi basati sull'informatizzazione dei *corpora* (ved. *Linguistica dei Corpora*, *Linguistica Computazionale*, *Studio del Linguaggio Naturale e Traduzione Automatica*) che hanno permesso di progredire negli studi fraseologici. Altresì, sono tanti gli studi e i congressi che ci informano riguardo i progressi dei metodi informatici innovativi basati in *corpora* applicati alla fraseologia monolingue, bilingue e plurilingue. Tuttavia, possiamo affermare che lo studio svolto nel volume in oggetto si può considerare di grande utilità giacché la materia trattata è innovativa ed è grande l'opportunità che offre agli studiosi interessati di poter ampliare i propri orizzonti.

A questo riguardo, dobbiamo dire che Antonio Rico-Sulayes possiamo considerarlo un grande esperto di fraseologia computazionale. Noi, che al contrario non abbiamo grandi conoscenze informatiche e quindi non vogliamo fornirvi affermazioni fuorvianti a sfavore di questo studio così approfondito, ci limitiamo a mettere in risalto quanto afferma l'autore nel capitolo IV circa

l'uso di sequenze di parole, chiamate 'n-grammi'. Sono strutture che hanno molte applicazioni nella linguistica computazionale e una presenza capillare nell'attribuzione autoriale. L'ampio uso delle sequenze di parole in linguistica informatica ha portato alla nascita di una fraseologia computazionale che si sta ritagliando sempre più spazio negli studi fraseologici, come si è osservato durante le giornate di studio EUROPHRAS 2015 e 2017. Anche se queste sequenze di parole inizialmente erano semplicemente combinazioni di due, tre o più parole nel testo, la frase computazionale mette in risalto le sequenze che rappresentano combinazioni con un certo grado di fissazione e senso idiomatico. Ad esempio, Rico-Sulayes fa largo uso, tra tante altre strutture linguistiche, dei giri fraseologici con funzioni grammaticali fissi, quali preposizioni o congiunzioni.

Quanto detto dimostra che la fraseologia è in un momento di grande fioritura volto ad influenzare le aree più diverse della linguistica, persino quelle spesso guidate da metodologie strettamente quantitative e che sono governate dai dati, come nel caso della linguistica computazionale.