

---

# CERLIS

## Journal

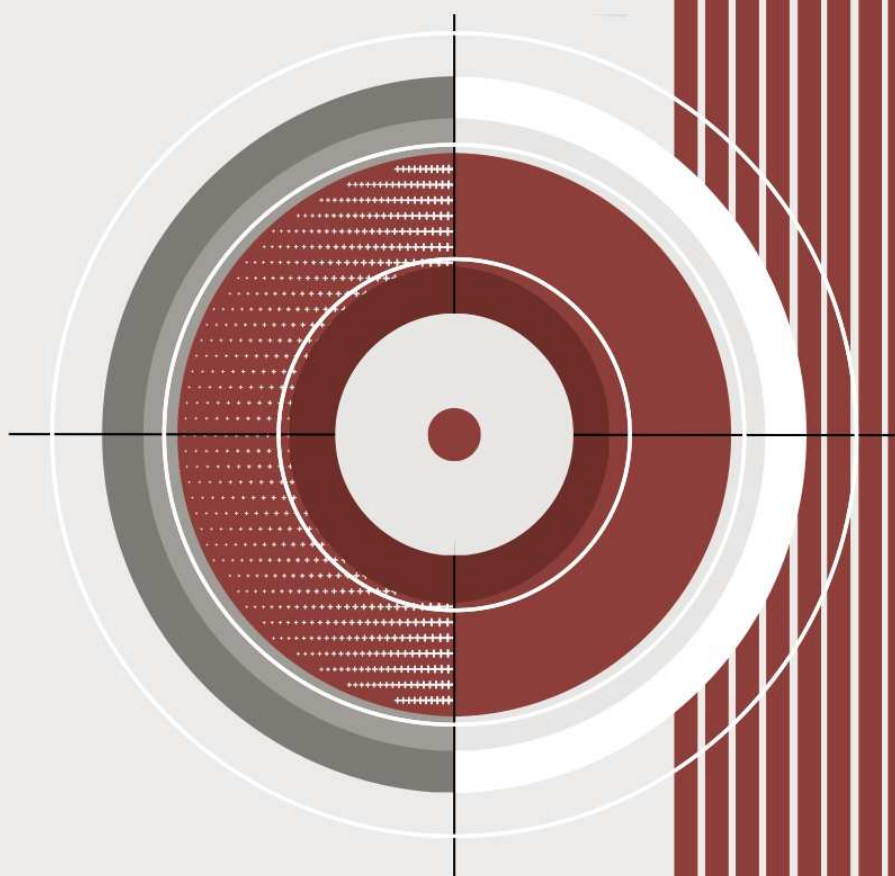
Studies on Specialized Languages  
and Discourse Analysis

*Special Issue*

*Advances in Computer-assisted Discourse Studies*

2

June 2025





# **The CERLIS Journal.**

## **Studies in Specialized Languages and Discourse Analysis**

---

Issue 2 | June 2025

Special Issue

Advances in Computer-assisted Discourse Studies

Exploring Discourse through Nooj

CERLIS Journal is an open-access, peer-reviewed publication dedicated to advancing research on specialised languages and discourse analysis. As a scholarly initiative of the Centro di Ricerca sui Linguaggi Specialistici (CERLIS) at the University of Bergamo, the journal provides an inclusive platform for exploring how language functions within diverse professional and academic communities. Welcoming contributions from both Italian and international scholars, CERLIS Journal embraces a wide range of theoretical and methodological approaches, including corpus linguistics, discourse analysis, genre studies, and translation studies.

CERLIS Journal is an open access journal, and article are published under CC BY-NC-ND 4.0 International license

CERLIS Journal adheres to the BOAI declaration: all users have the right to read, download, copy, distribute, print, search, or link to the full texts of these articles. This journal allows the Author(s) to hold the copyright and to retain publishing rights without restrictions and it does not charge APCs or submission charges. The PDF version of all the accepted articles and contributions can be accessed on-line and downloaded free of charge. The free on-line access to all contributions has been introduced to encourage a wider and more effective circulation of the ideas and findings originating from academic and scientific research.

Type: double-blind peer-reviewed journal of international scope

Frequency: published once a year

Language of contributions: French, English, German, Italian, Russian, Spanish

ISSN: 3103-7941

Licensed under a Creative Commons BY-NC-ND 4.0 International License

*Editor in Chief:* Stefania M. Maci (University of Bergamo, Italy)

*Deputy Editors:* Paul Baker (University of Lancaster, UK); José Santaemilia (Universitat de València, Spain); Jan Engberg (Aarhus University, Denmark)

*Scientific Board:* Giuseppe Balirano (University of Napoli l'Orientale, Italy); Begoña Bellés Fortuño (Universitat Jaume I, Castellón, Spain); Samuel Bennet (Uniwersytet im Adama Mickiewicza w Poznaniu, Poland); Silvia Bernardini (University of Bologna, Italy); Vijay Bhatia (City University of Hong Kong, HK); Ana Bocanegra-Valle (Universidad de Cádiz, Spain); Marina Bondi (University of Modena and Reggio Emilia, Italy); Silvia Bruti (University of Pisa, Italy); Piotr Cap (University of Lodz, Poland); Gabriella Carobbio (University of Bergamo, Italy); Paola Catenaccio (University of Milan, Italy); Larissa D'Angelo (University of Bergamo, Italy); Massimiliano Demata (University of Turin, Italy); Cécile Desoutter (University of Bergamo, Italy); Elena Di Giovanni (University of Macerata, Italy); Luisanna Fodde (University of Cagliari, Italy); Miguel Fuster-Márquez (Universitat de València, Spain); Giovanni Garofalo (University of Bergamo, Italy); Maurizio Gotti (Emeritus Professor University of Bergamo, Italy); Maria Grazia Guido (University of Salento, Italy); Cristina Hanganu-Bresch (Saint Joseph's University, Philadelphia, USA); Dorothee Heller (University of Bergamo, Italy); Violeta Jurkovic (University of Ljubljana, Slovenia); Marek Łukasik (Akademia Pomorska w Slupsku, Poland); Antonio Moreno Ortiz (Universidad de Málaga, Spain); Mark McGlashan (Birmingham City University, UK); Michele Sala (University of Bergamo, Italy); Elena Semino (University of Lancaster, UK); Maria Grazia Sindoni (University of Messina, Italy); Charlotte Taylor (University of Sussex, UK); Jerome Tessuto (University of Naples, Italy); Michael J. Zerbe (York College of Pennsylvania, USA)

*Editorial board:* Valentina Adami (University of Bergamo, Italy); Annelie Ädel (Dalarna University, Sweden); Patrizia Anesa (University of Bergamo, Italy); Ismael Arinas Pellón (Universidad Politécnica de Madrid, Spain); William Bromwich (University of Modena and Reggio Emilia, Italy); Maria Vittoria Calvi (University of Milan, Italy); Elena Carpi (University of Pisa, Italy); Daniela Cesiri (Ca' Foscari University of Venice, Italy); Luisa Chierichetti (University of Bergamo, Italy); Isabelle Clarke (University of Lancaster, UK); Stefania Consonni (University of Bergamo, Italy); Gabriel Cutrufello (York College of Pennsylvania, USA); Roberta Facchinetti (University of Verona, Italy); Eleonora Federici (University of Ferrara, Italy); Costas Gabrielatos (Edge Hill University, UK); Giuliana Garzone (IULM, Italy); Éva Jakusné Harnos (National University of Public Service, Budapest, Hungary); Karl-Gerhard Hempel (University of Salento, Italy); Sylvia Jaworska (University of Reading, UK); David Katan (University of Salento, Italy); Maria Kuteeva (Stockholm University, Sweden); Alessandra Lombardi (Università Cattolica del Sacro Cuore, Brescia, Italy); Danio Maldussi (University of Bergamo, Italy); Giovanna Mapelli (University of Milan, Italy); Sergio Maruenda-Bataller (Universitat de València, Spain); Anna Mauranen (University of Helsinki, Finland); Manuela Caterina Moroni (University of Bergamo, Italy); Giulia Adriana Pennisi (University of Palermo, Italy); Gina Poncini (Zayed University, Abu Dhabi, United Arab Emirates); Gianluca Pontrandolfo (University of Trieste, Italy); Ignazia Posadinu (University of Essex, UK); Chiara Preite (University of Milan, Italy); Sandrine Reboul-Touré (Université Sorbonne Nouvelle, France); Alessandra Rizzo (University of Palermo, Italy); Goranka Rocco (University of Ferrara, Italy); Tiziana Roncoroni (University of Bergamo, Italy); Elisa Rossi (Université Lumière Lyon2, France); Carmen Sancho Guinda (Universidad Politécnica de Madrid, Spain); Mike Scott (Aston University, UK); Lorella Sini (Università di Pisa, Italy); Cinzia Spinzi (University of Bergamo, Italy); Massimo Sturiale (University of Catania, Italy); Michela Tonti (University of Bergamo, Italy); Félix San Vicente (University of Bologna, Italy); Elodie Vargas (Université Grenoble Alpes); Sara Vecchiato (University of Udine, Italy)



## Contents

|                                                                                                            |    |
|------------------------------------------------------------------------------------------------------------|----|
| MARIO MONTELEONE                                                                                           |    |
| Introduction.....                                                                                          | 7  |
| BEATRICE FOLINO, MARIO MONTELEONE                                                                          |    |
| Nooj Grammars and Cybersecurity:<br>Possible Future Uses and Scenarios.....                                | 17 |
| CECILIA OLIVIERI, DAVIDE CARMINATI                                                                         |    |
| Unveiling Migrant Entrepreneurial Narratives:<br>A Discourse Analysis Approach for Inclusive Support.....  | 35 |
| AGATHE SEMLALI                                                                                             |    |
| Linguistic Analysis for Project Evaluation through NooJ: the EU<br>WeMED_NaTour Pilot Trip Assessment..... | 61 |
| MARIO MONTELEONE                                                                                           |    |
| Some Reflections on Italian Derivation and Computational<br>Morphology.....                                | 79 |



MARIO MONTELEONE<sup>i</sup>

## Introduction

### I

Defining with due consideration the genesis and scientific purposes, as well as the importance of the volume we are on the point of presenting here, is a difficult task. Indeed, the most complex aspect to manage is that the essays this volume includes deal with highly hybridized research methodologies, which associate the practices of a specific type of Natural Language Processing (NLP) with those of other extremely dissimilar disciplines, such as Sociology, Cyber Security and Italian Morphosyntax. Hence, one may wonder how three essays with such different contents may refer to a common analytical framework so to be publishable together, in the distinct volume we are here introducing.

First, for this courageous yet thoughtful, scientifically motivated, and significant choice, we must thank all the Editors and all the Members of the Scientific and Editorial Boards of the CERLIS Series. With their profound expertise and great intellectual predisposition towards new and innovative research paths, the Scholars who manage the Cerlis Series have enthusiastically embraced not only the hybridizations described here, but also the methodological framework from which they arise. Indeed, the following pages will introduce the readers to the two crucial elements that unite the three papers, and without which these hybridizations would not have been possible: Rule-Based NLP and the NooJ software environment.

Regarding the software environment just mentioned, the extensive references present within the essays of the volume will provide readers with the information useful to acquire knowledge about NooJ functionality, routines and potentialities, as well as about both its creator, Max Silberstein, and the prestigious scientific community that has developed around this software environment.

At the same time, we believe that Rule-Based NLP requires a more detailed and in-depth exposition, if only because it opposes, in its methodological, functional, historical-linguistic and epistemological approach, the NLP method most commonly used today, which is based on the probabilistic – better known as statistical-stochastic – analysis of human linguistic production.

As is well known, today the aforementioned method is fundamental to fields such as Machine Learning and Machine-Learning Algorithms, hence to the various tools and types of Artificial Intelligence, as for instance weak (or narrow) AI and strong (or general) AI, and to Deep Learning, Chain-of-Thought techniques, Artificial Inference and Artificial Neural Networks building. In these tools and disciplines, the management and development of natural language is entrusted to Large Language Models (LLMs), a type of artificial intelligence pre-trained on large amounts of textual data and which uses probabilistic occurrence models to produce texts in human language. The tasks these models are designed to perform, and in which an attempt is made to simulate human language production, are wide-ranging, including among others automatic translation and summarization, Question-Answering and Problem Solving.

The underlying architecture of LLMs is often based on neural network models, such as Transformers, which process sequential data, typically found in texts. The algorithms underlying these models are pre-trained by reading a very large number of words and word sequences. During this pre-training, the models produce statistical patterns that highlight which words tend most often to follow other words - hence the probabilistic approach. Unlike humans, when reading text data, these algorithms are therefore unable to understand the contextualized meaning of the words and word sequences they encounter. However, they can memorize them, weighting them statistically, also producing formal contextual abstractions, so as to attempt to predict, given a word, the most likely word to follow. After

this phase, to refine LLM capabilities and improve performance in a particular task or domain, pre-training results are subjected to a fine-tuning operation, that is, further training on automatically and probabilistically smaller datasets, specifically labeled and relating to one or more domains of human knowledge. Therefore, if, for example, a healthcare organization wants to use an LLM to assist doctors in generating patient reports from text notes, the LLM pre-training results would need to be optimized by integrating medical and healthcare-related terms.

We will not delve into a detailed qualitative assessment of the linguistic production of all the disciplines and tools just mentioned, including LLMs. Indeed, we are convinced that exist essayistic fields more pertinent for such an analysis. As for the current state, we entrust to the most attentive and curious readers a deeper exploring of Artificial Intelligence issues and limitations. These include, among others:

- The lack of both a reliable deductive logic and a *Weltanschauung* even modestly imitative of the human one;
- The methodological poverty of artificial inference which, being decontextualized from an interpretative point of view, is in fact unable to produce reliable deductions, inductions and abductions;
- The inability to review and correct errors autonomously, hence the ethical risks that arise from such an inability;
- Finally, the limitations of automatic text production - and here we are referring among others to incorrect translations or differential equations, inappropriate, dangerous, and biased linguistic content - in short, everything that, in the assessment of current Artificial Intelligence, is known as ‘hallucinations’.

Nowadays, several authoritative studies exist that recommend against placing excessive reliance on LLM linguistic productions, in the absence of an earlier, careful and thorough fact-checking procedure. Among these, we cite the 2021 book by Erik J. Larson *The Myth of Artificial Intelligence: Why Computers Can't Think the Way We Do* (Harvard University Press and Belknap Press). We cite also the AAAI 2025 Presidential Panel on the Future of AI Research, published in

March 2025 by the Association for the Advancement of Artificial Intelligence, and freely available on the Web. Here, however, what we feel is our duty is attempting to refute beliefs we consider erroneous, which yet are widely held and somewhat connected to the relationship between Artificial Intelligence and Natural Language.

The first of these beliefs is that NLP is only a specific field of Artificial Intelligence, and that it deals with enabling computers to understanding human language and performing interactions by means of such an understanding, using LLMs as key tools. The second belief is that natural language, as well as the world's languages, can be accurately described on a probabilistic basis. In the following pages, we will attempt to refute these beliefs, recalling and citing specific people, places, historical moments, and events. To avoid making the reading more strenuous, we will keep footnotes and extra-textual references to a minimum, knowing that any references to what we have cited will be easy to find on the Web.

## II

To argue against the first belief, we will need to touch briefly upon the History of General Linguistics. We will begin by citing Ferdinand de Saussure, whose reflections indirectly gave rise to the field of structural linguistics, whose illustrious representatives have included, to name just a few, Zellig S. Harris, Noam Chomsky, Lucien Tesnière, Maurice Gross, and Joseph Greenberg. Anyone who has had the opportunity to read even a single work by these scholars knows that they not only laid the foundation for the autonomous structuring of NLP, but also guided subsequent research for years.

Furthermore, the very beginnings of NLP research confirm its autonomy from Artificial Intelligence. Indeed, the first paper on the potential intelligence of machines was Alan Turing's article 'Computing Machinery and Intelligence', which in 1950 introduced the famous Turing Test. Besides, in 1955 John McCarthy was officially to coin and use for the first time the term 'Artificial Intelligence'. Later,

in 1956, McCarthy, along with Marvin Minsky, Claude Shannon, Nathaniel Rochester, and Terry Winograd, organized the Dartmouth Workshop<sup>1</sup>, during which, for the first time, the functions and goals of Artificial Intelligence were defined. For those who at the time were very young, less than thirty-year-old researchers, Artificial Intelligence should *proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves.* From this statement of intent, it is possible to deduce that the young researchers behind this initiative already clearly foresaw the challenges in making a machine simulate the use of human reasoning and language. As we know, these challenges remain unresolved today, while the functional goals of Artificial Intelligence, as defined at Dartmouth, have not yet been achieved. Moreover, here, as for the relationship between machines and natural language we feel the urge to stress the semantic difference between “to use” and “to understand”, which brings us back to underline the differences in self-awareness that distinguish human beings from machines.

We also need to stress that the first concrete NLP applications were developed long before the Dartmouth Workshop, thanks to the automatic cryptanalysis efforts conducted during World War II by the Bletchley Park research group. These researchers, led by the already famous Alan Turing, successfully developed automatic techniques and strategies for decrypting messages produced by the Nazis with the Enigma machine, designed by Arthur Scherbius. At the time, and for the first time, the researchers of Bletchley Park used and implemented the so-called ‘Bomba’, an electromechanical decryption device created from an earlier version of a special calculating machine, used by Polish counterintelligence, before the invasion of Poland, to decipher secret German encoded messages. In fact, in addition to Turing and numerous other cryptographers, including Gordon Welchman and Joan Clarke, the Bletchley Park team also relied on the contributions of Marian Rejewski, a Polish mathematician who had created the first version of

---

<sup>1</sup> [https://en.wikipedia.org/wiki/Dartmouth\\_workshop](https://en.wikipedia.org/wiki/Dartmouth_workshop)

the ‘Bomba’. Although there were no linguists in the Bletchley Park group, we can affirm that this was the first moment in History in which natural language was processed automatically, thus giving rise to NLP at a time when Artificial Intelligence had not even been theorized.

We also like to remind that that the concept of Machine Translation, thanks to Google Translate today well known to much of the world, was already being developed in the 1930s by Georges Artsrouni and Peter Troyanskij, who presented two innovative proposals regarding the first patents for translation machines. These proposals included both a pre-editing phase of the text, in which the editor organized the words in the source language to manage their logical form, and a post-editing phase, in which the editor arranged the final output. Between these two phases, the machine translated the text into the target language. These proposals remained virtually unknown until the outbreak of the Cold War in the 1950s, when Warren Weaver, mathematician and researcher at the Rockefeller Foundation, produced the first paper on Computer Assisted Machine Translation, entitled “The Translation Memorandum.” Shortly thereafter, in 1954, IBM introduced the 701 Electronic Data Processing Machine, also known as ‘The Brain’, IBM first commercial scientific computer. Thus, in January 1954, using ‘The Brain’, the Georgetown Experiment was held at IBM headquarters in New York, marking the first public demonstration of a machine translation system. During the experiment, 49 sentences were translated from Russian into English using a system capable of processing 250 vocabulary words. The system undoubtedly had significant quantitative limitations, as it could only translate texts word for word, without any contextual knowledge. However, at a global level, it helped stimulate public interest and research in the field of machine translation. Furthermore, it dramatically increased translation speed but inserted a new set of challenges. Indeed, human interpreters use experience and instincts to reflect context, idioms, humor and structure in their translations. In an effort to replicate such nuance, the IBM team designed a program that incorporated algorithms based on grammatical and semantic logic. The results were thrilling but far from perfect. In 1959, for instance, IBM Research News, the company’s monthly newsletter about scientific developments, described how ‘The Brain’ had begun to produce *a quality of baby talk that amazes and delights linguistic scholars.*

Around the same time that Machine Translation was born, the first operating systems also emerged, making it possible to improve translation speed, but not avoiding the occurring also of Machine Translation main limitations. Mathematician Yehoshua Bar-Hillel argued that fully automated translation would be possible only by struggling with a low level of quality in the final result. Furthermore, he believed that semantic ambiguity and syntax complexity were the two main obstacles to the development of fully automatic, high-quality translation.

Over the next decade, research continued, focusing in particular on the translation relationship between English and Russian, as well as the translation of technical and scientific documents. However, a turning point came in 1966 with the ALPAC report, commissioned by the US and delivered by the Automatic Language Processing Advisory Committee. The report dampened research enthusiasm in the field of Machine Translation, highlighting its limitations. In particular, it underlined the lack of progress and the marked differences, in terms of the low quality, compared to human translation. This led to a severe halt in funding for machine translation, which was deemed unreliable, plus extremely time-consuming and expensive.

Based on what we have highlighted, we can safely say that by 1966, NLP had already developed thirty years of research. Therefore, it is possible to conclude that NLP was already an autonomous discipline in historical periods when Artificial Intelligence, theorized in 1956, was still a fascinating hypothesis. Today, on the Web and elsewhere, it is possible to read, even from illustrious sources, the claim that NLP is only a branch of Artificial Intelligence. This is therefore an extremely inaccurate statement.

### III

We will now address the second belief we need to invalidate, namely the possibility of accurately describing natural language through probabilistic abstractions, using the already mentioned LLMs, that is to

say large data repositories. To use a perhaps not entirely elegant but certainly applicable comparison, we can say that from this perspective, LLMs, to model languages and speech, use the same approach as T9, namely the one introduced by Markov Chains, to suggest, certainly in more sophisticated terms, the occurrence of a word based on the one immediately preceding it.

As far as NLP is concerned, being probabilistic mechanisms Markov Chains present methodological limitations arising from the lack of linguistic tools specific for natural language analysis and production. Actually, we might argue that Statistics equips Markov Chains with nonspecific tools, i.e. with tools applicable to extremely dissimilar events, pertaining to different systems, all of them with specific idiosyncrasies that such tools cannot account for, due to their methodological nature. In our case, as for natural language, Markov Chains do not encompass formalized morph-syntactic rules. In this sense, this probabilistic approach to natural language is prone to produce almost always results that are not entirely reliable, in both text analysis and production phases.

Despite what we have just stated, our aim here is not to discredit Statistics and its analytical tools, which perform precisely the tasks for which they are designed. However, it is straightforward that from a methodological perspective, the Markovian, statistical and probabilistic approach to natural language lacks the depth necessary to understand and convey the morph-syntactic, logical-deductive, semantic, and inferential relationships that humans establish between words, sentences, and discourses whenever they utter language meant to be functionally communicative. This happens because probabilistic calculus, however refined, is not structured in such a way as to govern the aforementioned relationships, and can only observe their occurrences without however having or giving knowledge of their specific correlative and selective mechanisms.

We add to this that, as expressed by the magazine *Nature* in an online article dated December 16, 2024, “*probability probably doesn’t exist (but it is useful to act like it does)*”. Specifically, the author of the article, David Spiegelhalter, states in a reasonably understandable way that “[a]ttempts to put numbers on chance and uncertainty take us into the mathematical realm of probability, which today is used confidently in any number of fields. Open any science journal, for example, and you

*will find papers liberally sprinkled with  $P$  values, confidence intervals and possibly Bayesian posterior distributions, all of which are dependent on probability. And yet, any numerical probability, I will argue — whether in a scientific paper, as part of weather forecasts, predicting the outcome of a sports competition or quantifying a health risk — is not an objective property of the world, but a construction based on personal or collective judgements and (often doubtful) assumptions. Furthermore, in most circumstances, it is not even estimating some underlying ‘true’ quantity. Probability, indeed, can only rarely be said to ‘exist’ at all.”* In practice, translating this statement to natural language, it is possible to affirm that LLMs do not describe objective probabilistic models, but only sets of habits prevalent in linguistic concatenations, the deep dynamics of which, in any case, exist and are governed by elements outside both statistics and probability. This condition, therefore, offers us a way to demonstrate the coherence and validity of the linguistic analysis method we have adopted in this volume, which we previously have mentioned as ‘Rule-Based NLP’. To this end, we will make a strong and, perhaps, politically incorrect statement: in natural language analysis, probability does not appear to have an indispensable function. In truth, from a productive and concatenation perspective, probability is never applied in the creation of acceptable and, at the same time, grammatical linguistic content, and clashes with another aspect, which is that of possibility. The two quick examples that follows demonstrate how, thanks to morph-syntax, it is possible to predict the nature of relationships between words in terms of possibility and certainty, avoiding probability. For instance, in Italian, we know that a direct object will certainly, and not probably, follow or precede a transitive verb. Similarly, we know that a singular masculine adjective will certainly, and not probably, agree in gender and number with a singular masculine noun. To all this, we add that in the semantic economy of a given sentence, it is possible, and not probable, that a transitive verb deletes its direct object. In addition, it is possible, but not probable, that a masculine singular noun occurs not followed or preceded by a masculine singular adjective. This means that, in a given language, any concatenation of words is governed by morph-syntax, which encompasses possibilities and certainties, namely those governed by

rules on grammatical agreement and morph-syntactic dependence between words.

Hence, the question arises: why should we rely on the probabilistic study of natural language when, in fact, we have morph-syntax that defines and explains in detail the rules of linguistic concatenation? Moreover, this is a widespread condition not only in Italian, or Romance languages, but also in all the inflected languages of the world, as well as in those that retain case grammars. Furthermore, the usefulness of morph syntax compared to the probabilistic approach can be evaluated on a much more extensive basis. Here, we refer to the syntactic-semantic rules that direct and allow to predict with certainty, starting from a given predicate, the co-occurrence of certain morph-syntactic and semantic classes of words and the precise restrictions applicable in the selection of such words. An NLP method that formalizes these rules is, indeed, rule-based, and as the articles in this volume highlight, it offers the advantages of being flexible yet reliable, bottom-up, easily scalable, modular, highly context-sensitive, and iterative from a modelling perspective. Furthermore, it protects against any typos in texts, which can therefore be automatically recognized and corrected. The NLP method *par excellence* structured as we have just described is Lexicon-Grammar, developed by Maurice Gross in the 1960s. It represents the core structure of NooJ, and we are extremely pleased to recommend the readers of this volume to explore the Lexicon-Grammar method, for which they will find extensive information on the Web.

In conclusion, we would like to offer a final reassurance. While reading the three essays pages of this volume, the more theoretical and historical-narrative aspects exposed in this introduction will certainly be clearer and more comprehensible. As well, being those who have had the prestigious responsibility and honour of introducing the volume, we would like to thank once again all the Editors and Members of the Scientific and Editorial Boards of the CERLIS Series for their distinguished hospitality.

We wish everyone a pleasant and valuable reading.

BEATRICE FOLINO <sup>i</sup>  
MARIO MONTELEONE <sup>ii</sup>

## NooJ Grammars and Cybersecurity: Possible Future Uses and Scenarios

### Abstract

According to the Italian Observatory of Digital Innovation, "Cybersecurity is the main digital investment priority in Italy, also thanks to the push of the PNRR and the guidance of the National Cybersecurity Agency. However, the relationship between spending on cybersecurity and GDP is 0.10%, still far from the other G7 countries: 67% of companies in Italy note an increase in attack attempts, 61% have increased their budget for IT security. In 53% of companies, there is a formalized Chief Information Security Officer. In 80% structured training plans on attack risks. Only a third of companies have methodologies for financial quantification of cyber risk. Cyber-attacks are continually increasing, with 1,141 serious incidents detected by Clusit in the first half of 2022 alone, +8.4% compared to the same period in 2021, and threats are increasingly also affecting critical infrastructures. In this context, 67% of companies note an increase in attack attempts and 14% have suffered tangible consequences following cyber incidents, such as service interruptions, delays in process operations or reputational damage.

Keywords: NooJ, NooJ Grammars, NooJ Finite-State Automata and Transducers, Cybersecurity Algorithms, Unix Network Security Model, Apache Web Server Security

---

i      Epicode Italia, Rome – Milan – Berlin – [beatricefolino@outlook.it](mailto:beatricefolino@outlook.it)  
ii     Università degli Studi di Salerno, Italy – [mmonteleone@unisa.it](mailto:mmonteleone@unisa.it); X-23, Italy -  
      [mario.monteleone@x-23.org](mailto:mario.monteleone@x-23.org)

## 1. The Growing Threat of Cyber-Attacks

According to Assintel<sup>1</sup>, cyber-attacks steadily increased during 2023: compared to the previous year, there was an impressive increase of 184%, with a total of 7,068 attacks identified and classified in the world. In 61% of cases, the attacks came from the Dark Web, without any evidence on the Clear Web. These are some of the data contained in the first annual report on the evolution of cybersecurity created by Assintel - Confcommercio through its Cyber Think Tank. According to data collected by the Assintel working group, geographically cyber-attacks have increased to 50% in America and 27% in Europe. As far as Italy is concerned, in the first half of 2023 there was a +85.7% increase in attacks compared to the previous quarter. In addition, SMEs, especially small micro-businesses, have proven to be hackers' favorite targets. It ensues that in Italy attention is growing for cybersecurity, which in 2023 is confirmed as the main digital investment priority among companies, both large and SMEs. As many as 61% of organizations over 250 employees have decided to increase their budget for cybersecurity activities in the last 12 months<sup>2</sup>. Overall, in 2022, the Italian cybersecu-

---

<sup>1</sup> <https://www.assintel.it/wp-content/uploads/2024/04/Assintel-Cyber-Report-2023.pdf>. Assintel's Cyber Report also highlights how attack techniques present an increase in the use of Malware, reaching 70% of total attacks in 2023. Furthermore, the impacts of cyber-attacks worry experts: 91% of attacks in 2023 are classified as serious or very serious and those with critical impacts represent 24%, highlighting significant repercussions in economic, legal and reputational terms for the victims. The manufacturing sector was hit hardest by attacks, rising from 5% to 16% of total attacks in 2023, followed by professional/scientific/technical, ICT, healthcare, and financial/insurance sectors. The most used techniques for cyber-attacks are malware, which has reached 70% of the total attacks, followed by the use of vulnerabilities and unknown techniques. Nearly a quarter of attacks had critical impacts, while 67% had severe impacts. This indicates a significant increase in attacks with catastrophic economic, legal or reputational consequences for victims.

<sup>2</sup> However, despite the increase in spending on cybersecurity (which represents 0.12% of Italian Gross Domestic Product), our country is still in last place in the G7. The ranking is led by the United States and the United Kingdom, with a ratio of 0.31%. For France and Germany, the ratio is 0.19% and 0.18% re-

rity market has reached a value of 1.86 billion euros, with an exceptional acceleration of +18% compared to 2021. The ratio between spending on cybersecurity and GDP in Italy stands at 0.10%, slightly up compared to 0.08% the previous year.

### *1.1 Main Cyber-Attack Tools and Methods*

At present, Cybersecurity routines mostly use encryption algorithms. According to Rootstrap [2], “An encryption algorithm is a mathematical procedure or set of rules that can either perform or determine how data is to be encrypted. An algorithm converts information into meaningless ciphertext and requires a key to transform the data into its original form. Simply put, the best encryption algorithms are programs that easily allow a sender to scramble their data and then permit a receiver to unscramble it on the other end. [...] Many different encryption algorithms rely on various methods of encryption. The following are the best encryption algorithms worth looking into or employing [...]: 3DES Algorithm, AES Algorithm, RSA Algorithm, ECC Algorithm, BlowFish and

---

spectively. The 2024 report from the Cybersecurity & Data Protection Observatory of the School of Management of the Polytechnic of Milan highlights that 62% of large organizations have increased spending on cybersecurity. This increase was driven by the introduction of new tools (68%), greater attention from part of the boards of directors (62%) and by the need to adapt to regulations (43%). However, small businesses struggle to turn this interest into concrete investments due to limited resources and a lack of market offerings that meet their specific needs. Gabriele Faggioli, Scientific Manager of the Cybersecurity & Data Protection Observatory, underlines the importance of balancing technological investments and human capital to reduce the gap between Italy and other countries. On the one hand – he explains – it is essential to exploit the potential of innovative technologies such as artificial intelligence. On the other hand, it is essential not to underestimate the human component, insisting on training and raising awareness of workers. Alessandro Piva, Director of the Cybersecurity & Data Protection Observatory, highlights that Artificial Intelligence (AI) can be both a weapon in the hands of cybercriminals and a defence tool for companies. 56% of organizations use AI to improve the protection of information assets. The Observatory's research also highlights an increase in supply chain attacks and an increasingly close link between cyber-attacks and technologies that use artificial intelligence algorithms. Finally, employee training and awareness represent one of the main priorities for action in the cybersecurity field for 71% of large companies.

TwoFish. It is important to note that these algorithms primarily rely upon the properties of mathematics.”

### 1.2 Aims of this Study

To achieve the aims of this study, we rely on what recently stated by NVidia researchers [3], that is the fact that “the role of Natural Language Processing [...] in the cybersecurity-world starts to gain traction. So what may be the role and applications of NLP in cybersecurity? [...] people started communicating with machines through constructed language (e.g. programming languages, etc.) but are increasingly using the natural language to do it (e.g. chatbots, virtual assistants, etc.). [Therefore] here are some NLP tasks that may be considered in creating threat or attack tools, or in defending humans and machines:

- Language Analytics: language modelling, sentiment analysis, text classification, named-entity recognition, natural language inference, relation extraction, semantic parsing, co-reference resolution, entity linking, relational reasoning, semantic composition, language identification and translation, entity and information extraction, intent detection and classification, stance and fake news detection, rumor detection, hate speech detection, clickbait detection, abuse detection.
- Language Generators: question-answering systems, text and dialogues generation, text summarization, slot filling for knowledge base population tasks, scripts and programming code generation.

Hence, in the following pages, we will verify in and how it is possible to insert (parts of) Cybersecurity algorithms instructions into specific NooJ grammars, in order to use NooJ as a device to detect, highlight and challenge cyber-attacks automatically and in real time. Ultimately, we will also try to outline the structure of a “standardized” automatic NLP system to support the management of similar attacks.

## 2. Most Used Cyber-Attacks Methods

We can subdivide cyber-attacks methods into two macro-categories<sup>3</sup>: syntactic attacks, which are "direct" and based on the (involuntary and hidden) dissemination and use of malicious software; semantic attacks, which are "indirect" and include modification of correct information and dissemination of false information. For the sake of brevity, we will give below a brief overview only of the most frequent cyber-attacks of the first group.

**Malware:** deriving from the contraction of "malicious" and "software", the term literally means "evil program" and indicates any computer program capable of damaging the functioning and security of the operating system. It is transmitted across the Internet, often through email or simple browsing, and the most common varieties include viruses, Trojan horses, keyloggers, worms and backdoors. There exist "polymorphic" malware, which continuously change shape, and "metamorphic" malware, which completely alter their code. Difficult to detect, they do not damage the physical hardware of a system, nor the network, but are capable of stealing, encrypting or deleting data, altering or compromising the fundamental functions of a computer and spying on users' activities, without them realizing it or providing any authorization in this regard.

**Phishing:** one of the most used cyber-attacks, used also to spread serious malware, such as ransomware. This cyber-attack "fishes" for users' financial data and password. This is a type of fraud carried out on the Internet, through which victims are deceived to provide personal and confidential information, financial data or credit card access codes, posing as a financial institution or other reliable body. Usually, in the message sent to the user, a link is indicated which only apparently refers to the website of the credit institution or the service for which one has registered. Ultimately, when entered by users, all confidential data are "stolen" (i.e. recorded and reused) by attackers.

**Man-in-the-middle attack:** this type of attack occurs when a third party secretly retransmits or alters communication between two other parties who believe they are communicating directly with each other. In the most common example, the attacker uses a Wi-Fi network

---

<sup>3</sup> <https://universeit.blog/attacchi-informatici/>.

to intercept users' communications, attacking the router connection apparently without flaws or exploiting a weak point in the router setup. A recent variant is the "man-in-the-browser" attack, in which cybercriminals install malware on victims' computers, recording the data exchanged between the browser and the target sites where the malware code has been inserted. This type of attack allows to hit multiple people at the same time and can be carried out remotely.

**DoS attack – Denial-of-Service:** it refers to a malfunction caused by an attack in which the resources of a computer system providing a service to customers are deliberately depleted, until it is no longer able to function. This attack causes a "denial of service", overloading, with a myriad of requests, the network connections of a system used for the exchange of external data. When the quantity of requests exceeds the capacity limit, the system slows down or collapses. Typical targets of DoS attacks are online shopping sites, online casinos, and any company and organization that provides online services.

**SQL injection:** in this attack technique, attacker targets the vulnerabilities typical of those databases that use the SQL language for data entry. In particular, it targets those user inputs not filtered correctly and characterized by the presence of some metacharacters, including, for example, the double hyphen, quotation marks and semicolons. For the SQL interpreter, all these metacharacters allow the external modification of commands. The ideal target for this cyber-attack are programs that have old interfaces and in which the inputs are not filtered correctly. In this way, by means of a precise use of function characters, any user without authorization can inject other SQL commands and end up tampering with the database.

**Zero-day cyber attacks:** they represent the latest generation of cyber threats and are referred to as Advanced Persistent Threats (APT). It is very difficult to understand where they come from, even if generally it is a cybercriminal (or a group) who has discovered certain vulnerabilities and has begun to exploit them, including those within browser services and applications by email. The names "zero-day" - or "zero days" - means that, since these are newly discovered security flaws, "zero" is the time available to developers to fix them before they can be exploited. Besides, there is a distinction made between "zero day vulnerabilities", which refer to holes in the protection of software installed on a browser or an application, and "zero day exploits", an attack

that exploits these vulnerabilities to install malicious software on a device.

**DNS Tunnelling:** attackers hide data within the Domain Name System (DNS<sup>4</sup>) queries and send such data to transfer or activate malware inside a server. The most common attacks are against command and control servers: once an internal device has been compromised, the attacker will maintain contact with that device to execute commands.

**Brute-force attack:** Brute Force Attack is a method used by cybercriminals to crack account passwords and discover login credentials. This type of attack relies on a dictionary of common words and passwords that is used to attempt to discover the victims' password. After exhausting all the terms in the dictionary, cybercriminals move on to more sophisticated techniques, using combinations of characters, until a match is found. It can take thousands of attempts to crack a password; this is why during a brute force attack automatic tools are used that allow a large number of attempts to be made in a short time. Brute force attacks can be launched against an application or against the encrypted value or hash of a password. Web applications are usually equipped with security mechanisms that block automatic login attempts typical of brute force attacks, so it is much more likely that cybercriminals use this type of attack directly with stolen passwords. If the attack is launched against an application, automatic software will be used to try a list of usernames and passwords until a match is found, to access to victims' account. A more common type of brute force attack is to guess the user's password from the encrypted value or hash of the password. To crack an encrypted password you need a private key. If an attacker came into possession of this private key, he would be able to decipher the password, or use automatic tools to try to decipher the value of the key. Passwords are normally saved in a hash version, one-way and without the possibility of deciphering. Therefore, cybercriminals use a dictionary of potential passwords, hash them, and if the value matches the hash of the stolen password. Brute Force Attacks are used to access illegally bank accounts, or steal personally identifiable information. The damage resulting from a brute force attack depends on the privilege level of the compromised account, and the type of application to which

---

<sup>4</sup> DNS is the protocol that keeps the network functioning, mapping domain names and matching them to their IP addresses.

that account is related. Some additional actions that cybercriminals can take following a successful brute force attack include:

- Send messages to the victim's colleagues or other users to encourage them to click on phishing links or open malicious attachments.
- Install malware on your system or network infrastructure. If an administrator device is infected, the attacker can steal high-level credentials.
- Sending messages to customers in an attempt to damage the victim's reputation.
- Hijack server processes to install malicious applications used to intercept outgoing and incoming traffic.
- Install adware on systems and applications, to earn revenue from advertisements.
- Redirect user traffic to a server controlled by criminals.

Brute force attacks are usually automatic<sup>5</sup>. A real person can only test a few passwords per minute, while a computer is capable of testing hundreds or thousands of combinations per minute, also depending on the

---

<sup>5</sup> Examples of password brute force attack programs are:  
**Aircrack-ng** (<https://en.wikipedia.org/wiki/Aircrack-ng>): a software suite for analysing and hacking Wi-Fi networks. Aircrack-ng functionality includes monitoring through packet capture and export of data to text files; attacking through deauthentication or fake access points.  
**John the Ripper** ([https://en.wikipedia.org/wiki/John\\_the\\_Ripper](https://en.wikipedia.org/wiki/John_the_Ripper)): dictionary attacks, a type of attack that exploits the search for words present in dictionaries or common language phrases.  
**L0phtCrack** (<https://en.wikipedia.org/wiki/L0phtCrack>): a password auditing and recovery application originally produced by Mudge from L0pht Heavy Industries. It is used to test password strength and sometimes to recover lost Microsoft Windows passwords, by using dictionary, brute-force, hybrid attacks, and rainbow tables.  
**Hashcat** (<https://en.wikipedia.org/wiki/Hashcat>): a password recovery tool. Until 2015, it had a proprietary code base, later it was released as open source software. Versions are available for Linux, MacOS, and Windows. Examples of hashcat-supported hashing algorithms are LM hashes, MD4, MD5, SHA-family and Unix Crypt formats as well as algorithms used in MySQL and Cisco PIX.

connection speed. Cybercriminals leverage automation to launch their brute force attacks. It is also common to see them using their own custom scripts, developed in their favorite programming languages, such as Python. In addition to password cracking tools, criminals also use vulnerability-scanning tools to identify outdated software and discover information about the targeted application.

### **APT Cyber-Attacks**

Considering the damage level they produce, APT attacks deserve a more detailed description. Behind the acronym APT<sup>6</sup> stands for Advanced Persistent Threats and their danger lies in the fact that no institution of any level is immune<sup>7</sup>. This phenomenon affects government

---

**Dave Grohl** (<https://en.wikipedia.org/wiki/DaveGrohl>): brute-force password cracker for MacOS. It was originally created in 2010 as a password hash extractor but has since evolved into a standalone or distributed password cracker. Dave Grohl supports all standard MacOS X user password hashes (MD4, SHA-512 and PBKDF2) used since OS X Lion. It can also extract them formatted for other popular password crackers like John the Ripper. The latest stable release is designed specifically for Mac OS X Lion and Mountain Lion.

**Ncrack** (<https://nmap.org/ncrack/>): a high-speed network authentication-cracking tool. It was built to help companies secure their networks by proactively testing all their hosts and networking devices for poor passwords. Security professionals also rely on Ncrack when auditing their clients. Ncrack was designed using a modular approach, a command-line syntax similar to Nmap (an open-source utility used for network scanning and security auditing), and a dynamic engine that can adapt its behaviour based on network feedback. It allows for rapid, yet reliable large-scale auditing of multiple hosts. Ncrack features include a very flexible interface granting the user full control of network operations, allowing for very sophisticated bruteforcing attacks, timing templates for ease of use, runtime interaction similar to Nmap and many more. Protocols supported include SSH, RDP, FTP, Telnet, HTTP(S), Wordpress, POP3(S), IMAP, CVS, SMB, VNC, SIP, Redis, PostgreSQL, MQTT, MySQL, MSSQL, MongoDB, Cassandra, WinRM, OWA, and DICOM.

<sup>6</sup> See <https://universeit.blog/apt-advanced-persistent-threats/>.

<sup>7</sup> Some examples: the Stuxnet APT knocked out Iran's nuclear program; Deep Panda attacked the US government's Office of Personnel Management, compromising more than 4 million US personnel records. APT34 targeted government organizations and financial, energy, chemical and telecommunications companies in the Middle East. The list of attacks is long and constantly growing, like the volume of spending to defend oneself: according to Statista, the global market for advanced protection tools from persistent threats worldwide will reach an estimated value of 10.6 billion in 2024 of dollars per year.

bodies and large companies, victims of Advanced Persistent Threats. In recent years, the number of reported cases related to APTs has increased significantly and the current moment, characterized by the Covid-19 pandemic, has generated the right scenario for the launch of new attacks. APTs are advanced persistent threats or complex and selective cyber-attacks, which, through unauthorized access to information and communication systems, are capable of reading confidential data or causing damage to a company, an industry or a government organization. The NIST - National Institute of Standards and Technology – defines the APTs as an adversary that possesses sophisticated levels of expertise and significant resources that allow it to achieve its objectives using both cyber and physical attack vectors. They pursue their goals repeatedly and over a pro-longed period (months to even years), adapting to defenders' efforts to resist them, and are determined to maintain the level of interaction necessary to carry out their tasks. APTs, duration over time and the difficulty of identification are the two peculiar elements that distinguish them from other cyber threats. If most malware performs a quick and malicious attack, APTs use more strategic and stealthy approaches. Attackers enter through traditional malware such as Trojans or phishing, but then cover their tracks while planting their attack software throughout the network. APTs are usually supported by nations or very large organizations. There are specific characteristics that indicate the existence of an APT attack. These signs include:

- **Actors:** attacks are usually carried out by people or groups pursuing a specific mission. These actors are often supported by nation states or corporate-backed organizations. As for objectives, however, they are aimed at undermining the target capabilities or gathering information over a prolonged period.
- **Timing:** attacks focus on ensuring that attackers can gain access and maintain access for a significant period of time. Among other things, attackers repeatedly return to a system they have infiltrated.
- **Resources:** APT attacks require significant resources to plan and execute.
- **Methods:** APTs often employ sophisticated techniques that require high security skills.

- **Origin:** APTs can come from a variety of places and can occur during an attack designed to distract security teams. Attackers often take the time to map comprehensively a system's weaknesses before choosing an entry point.
- **Value:** it can refer to the size of the target or the size of the attack operations. Large organizations are often the preferred victims of APTs more frequently than small ones<sup>8</sup>.

---

<sup>8</sup> An APT attack may have seven phases (structure developed by Mandiant Consulting, a company specialized in computer network security):

- 1) Initial reconnaissance: it is the first phase, during which the cyber attacker conducts research on a target, identifies it and determines its attack methodology. Scrolls the web looking for potential email addresses or social media accounts (Facebook, Twitter or LinkedIn etc.) Once the ideal contact information to target with spear-phishing is found, the attacker successfully executes the malicious code on one or more systems.
- 2) Initial compromise: The attacker successfully executes malicious code on one or more systems. This most likely occurs through social engineering (most often spear phishing), exploiting a vulnerability on an Internetfacing system, or by any other means necessary.
- 3) Establish a foothold: The attacker ensures that he maintains continuous control over a recently compromised system. This happens immediately after the initial compromise. Typically, it establishes a foothold by installing a backdoor (method of bypassing normal authentication in a computer system or cryptosystem) or by downloading additional utilities or malware onto the victim's system.
- 4) Privilege escalation: It is the phase in which the cyber attacker gains greater access to systems and data. It identifies the means to deliver the threat to the victim's computer. The three most popular forms of delivery are email, website and a USB stick. Now cloud delivery has also evolved.
- 5) Manipulation (Internal reconnaissance): The attacker explores the victim's environment to gain a better understanding of the organization/company, identifying key resources and people, and to determine where an organization stores information of interest. This is the phase in which the intruder triggers the execution of the threat. For example, in email delivery, attachments such as PDFs or Word documents are quite common. Then the intruder exploits vulnerabilities in these files to trigger the execution of malicious code.
- 6) Implementation (Lateral Movement): Lateral movement refers to the approach used by attackers to systematically traverse a network to access or damage valuable assets or data. To do this, it uses tools and methodologies to gain access and privileges, which allow it to move laterally across applications and devices in a network to isolate targets, map the system, and ultimately arrive at high-value targets. Cyber criminal gets easy access into targeted victim's system

Thus the attacker completes the mission and achieves his goal. Often this means stealing intellectual property, financial data, mergers and acquisitions information, or personally identifiable information. Once the mission is completed, most targeted attackers do not leave the environment, but retain access in case of a new operation. How to protect yourself from APTs: APTs are complex threats that must be addressed with equal strategy and competence. First of all, a broad-spectrum vision is needed that considers all parts of the organization even before the tools necessary to combat and eliminate the threat. As threats are advanced and persistent, a defence-in-depth strategy is needed to constantly monitor networks and security controls. We need an interdisciplinary approach, which refers to the criteria of multilevel defence. To discover suspicious events, proactive and real-time tools are needed such as monitoring processes, file and login operations, threat intelligence information that can help recognize typical patterns and internet addresses of cyber-attacks. One of the cornerstones for dealing with advanced and persistent threats is the Vulnerability Assessment. This is a security analysis whose objective is to identify all vulnerabilities in systems and applications. It does this by evaluating the potential damage that the possible "attacker" can inflict on the production unit through highly automated tools in a first phase, and then making use of the skills of highly qualified personnel who, at a second moment, integrates and verifies the results through meticulous manual activity. Today, artificial intelligence, in particular machine learning, is invaluable in addressing and solving cybersecurity problems such as APTs and detecting them. Some of the methods and techniques that researchers have used are directly related to ML. Prevention measures require greater analysis and response capacity in the shortest possible time, due to the large volume

---

by distributing malicious Trojan. So it uses its access to move from one system to another within the compromised environment.

7) Maintaining presence (Control): This phase is characterized by the cyber intruder's ability to guarantee control over the access channels and credentials acquired in the previous phases. ensures continuous access to the environment. Common methods to maintain a presence include installing multiple variants of malware backdoors or fraudulently logging into remote access services such as the company's Virtual Private Network.

of data and the rapid evolution of current threats. For this reason, automated tools have been created to assist cybersecurity administrators. Machine Learning techniques are a useful tool in the field of cybersecurity. For example, models of network traffic behaviour can be created to detect anomalous activity, reduce the number of false positives on alerts and detect threats in real time. However, machine learning can be used to create attacks on, for example, fraudulent email sending or password cracking software. ML applications in cyber-security come into play in three specific operations: detection, protection and prediction. In the first case we mean those tools that allow the detection of anomalous behaviour to generate real-time alerts and to facilitate the decision-making process. On the security side: ML applications can detect vulnerabilities to install automatically security fixes. Then there are techniques and algorithms dedicated to predicting attacks and developing anti-malware techniques.

### 3. The Nooj Shield against Cyber-Attacks

We have already mentioned in 1.2 the role NLP can have in Cybersecurity. Therefore, in the following pages, we will see how it is possible to use NooJ as a shield against cyber-attacks, more specifically during a brute-force attack. We will exemplify a NooJ session in which a brute-force text is processed to locate and retrieve malicious codes. In this regard, it is important to underline that cyber-attacks always present themselves as code texts, and that public libraries<sup>9</sup> of such code texts have long existed on the Web, hence can be viewed downloaded, and used to build NooJ electronic dictionaries, as well as transcribed in the graphs of NooJ grammars. Besides, it is worth remembering that the

---

<sup>9</sup> For some examples of code, see:  
<https://gist.github.com/raphaelvallat/646bd1675f2dadff09c50ebc85f298b8>;  
<https://github.com/danielmiessler/SecLists/blob/master/Passwords/probable-v2-top12000.txt>;  
<https://github.com/dominictarr/random-name/blob/master/middle-names.txt>.

NooJ Apply routine allows safeguarding and reapplying complete NooJ work sessions.

### 3.1 How to Use NooJ to Block a Brute-Force Attack

In case of a Brute Force Attack done With Python, NooJ can be installed as an active feature of the OS core on the central server. It works in a loop, using an electronic dictionary of possible virus-code strings to analyze all “anomalous” messages that the central server receives. For each anomalous message, the server activity is put on standby (it does not communicate with the rest of the company network). Each anomalous message is physically printed on paper, to have concrete data through which to recognize the cyberattack, in the event of interruption of the central server activities. In case of a concrete and certified attack:

1. Via a warning from NooJ, the OS shuts down the company central server;
2. (If not subject to attack) A backup server automatically turns on;
3. Via the backup server, the company network is reactivated to allow normal work activities;
4. The central server is physically disconnected from the company network;
5. The central server is turned back on for virus-cleaning procedures.

```
import requests
from termcolor import colored

url = input('[+] Enter Page URL: ')
username = input('[+] Enter Username For The Account To Bruteforce: ')
password_file = input('[+] Enter Password File To Use: ')
login_failed_string = input('[+] Enter String That Occurs When Login Fails: ')
cookie_value = input('Enter Cookie Value(Optional): ')

def cracking(username,url):
    for password in passwords:
        password = password.strip()
        print(colored(('Trying: ' + password), 'red'))
        data = {'username':username,'password':password,'Login':'submit'}
        if cookie_value != '':
```

```

        response = requests.get(url,
params={'username':username,'password':password,'Login':'Login'}, cookies
= {'Cookie': cookie_value})
    else:
        response = requests.post(url, data=data)
        if login_failed_string in response.content.decode():
            pass
        else:
            print(colored('[+] Found Username: ==> ' +
username), 'green'))
            print(colored('[+] Found Password: ==> ' + pass-
word), 'green'))
            exit()

with open(password_file, 'r') as passwords:
    cracking(username,url)

print('[!!!] Password Not In List')

```

Table 1. Example of Text of Brute-Force Attack Code.

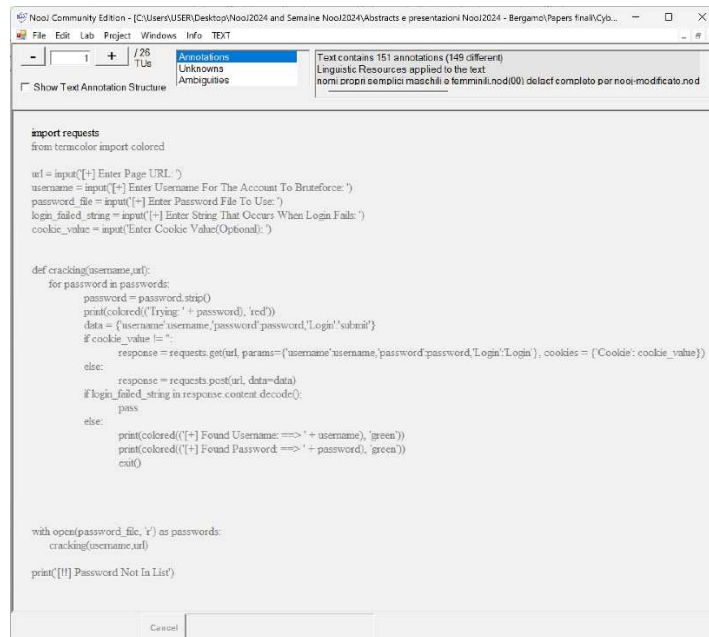


Fig. 1. Example of Brute Force Attack Code Text processed with NooJ

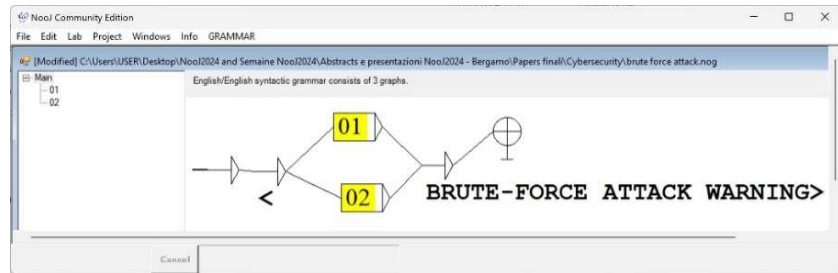


Fig. 2. A NooJ Finite-State Transducer Grammar for Brute-Force Attack Warning – Main Graph.

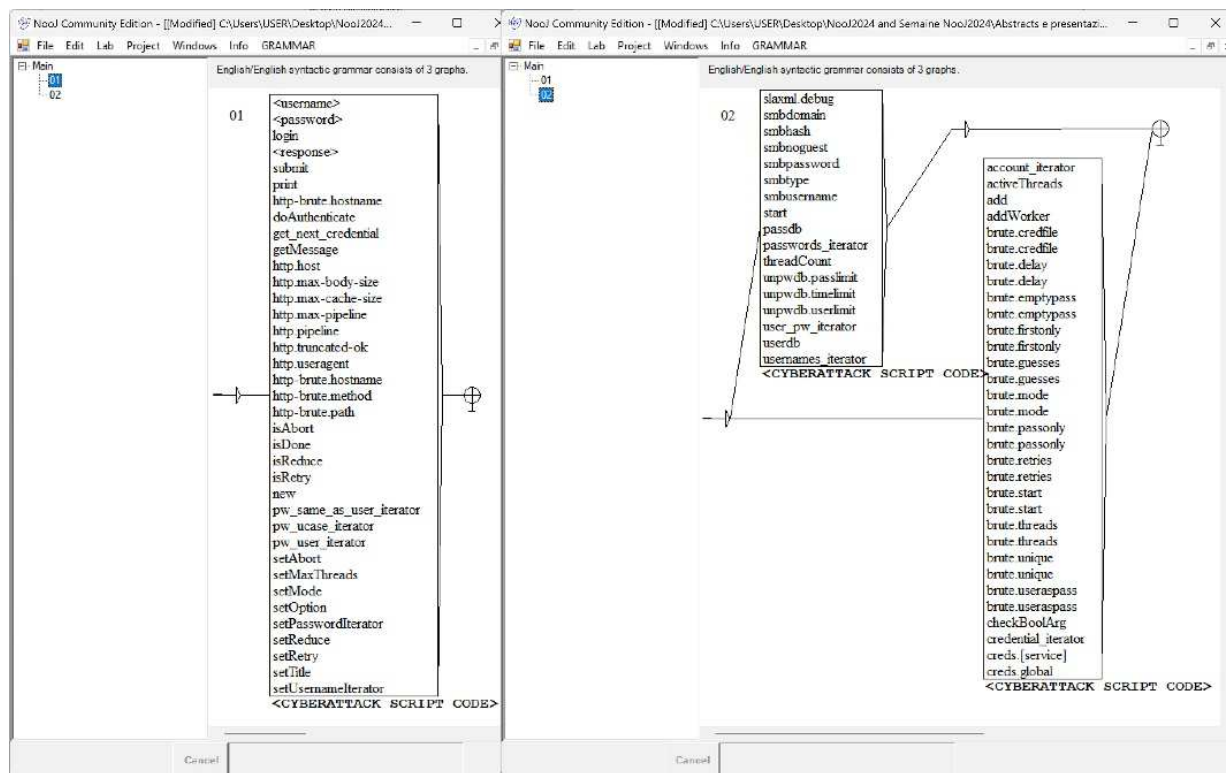


Fig. 3. A NooJ Grammar Finite-State Transducer for Brute-Force Attack Warning – Subgraphs 01 and 02.

## 4. Conclusions and Further Research

Our shield procedure carried out via NooJ against a Brute Force cyber-attacks has the quality of being almost completely automatic, with the exception of the creation of grammars and dictionaries. It is also plausible that similar procedures could be created and applied for many other types of attacks, especially given the fact that in their variety, they present themselves in the form of code texts. Our shield procedure carried out via NooJ against a Brute Force cyber-attacks has the quality of being almost totally automatic, with the exception of the creation of grammars and dictionaries. It is also plausible that similar procedures can be created and applied for many other types of attacks, especially given the fact that in their variety, they present themselves in the form of code texts. All this confirms, among other things, the extreme application eclecticism of NooJ, which will, in the future, give the possibility of creating a complex and effective shield system against cyber-attacks of various kinds.

## References

- [1] Web page, <https://www.osservatori.net/it/ricerche/comunicati-stampa/cybersecurity-italia-mercato-crescita>. Last accessed: February 8, 2024.
- [2] Web Page, <https://www.rootstrap.com/blog/the-best-encryption-algorithms-for-cyber-security>. Last accessed: February 13, 2024.
- [3] Web page, <https://developer.nvidia.com/blog/changing-cyber-security-with-natural-language-processing/>. Last accessed: February 8, 2024.
- [4] Silberztein, M.: The NooJ Manual. NooJ Association Web Page, <http://www.nooj-association.org>. Last accessed: 2024/19/09.
- [5] Silberztein, M. Formalizing Natural Languages: The NooJ Approach. ISTE Ltd and Wiley, London (2016).



CECILIA OLIVIERI  
DAVIDE CARMINATI

## Unveiling Migrant Entrepreneurial Narratives: A Discourse Analysis Approach for Inclusive Support

### Abstract

Migrant entrepreneurs in Europe constitute a multifaceted group, represented by complex identities considering multiple generations, places of origin, demographic characteristics, and ambitions. The exploration of this target group is a crucial endeavor, given the unique challenges and opportunities it faces exploring unfamiliar economic landscapes. Despite its sociological and political relevance, approximative narratives on migrant entrepreneurship are widespread. Support organizations in entrepreneurship training often portray migrants as vulnerable social groups and bearers of basic needs only. These narratives homologate their perception, preventing a diversified support system. Self-perception data directly collected from migrants offers a valuable alternative to these top-down narratives, providing authentic perspectives on their socio-economic contexts, needs, aspirations. Understanding these elements is crucial for developing appropriate support pathways. This study implements an innovative discourse analysis model using co-occurrence analysis through NooJ formal and semantic grammars to examine migrant entrepreneur interviews. By identifying core themes that hold significance for the entrepreneurs and their interconnections, the research strives to underlying beliefs and inclinations as entrepreneurs narrate their entrepreneurial journeys. By uncovering these connections, this work aims to provide a deep understanding of the speakers' needs, skills, and aspirations, informing effective incubation pathways for a more inclusive, empowering, and personally attuned support systems.

Keywords: NooJ, Discourse Analysis, Migrant Entrepreneurs, Tailor-Made Support Models, Incubation Programs.

## 1. Introduction

Global migrations present challenges and opportunities for both the hosting countries and the migrants. The recurrent narrative from the media stresses that labour integration is central to the process of inclusion. Still, despite this position, concretely observable facts show a different perspective where migrants are undervalued, have more difficulties in accessing the job market, experience discrimination, have lower salaries, and are more likely to be placed in the unregulated and informal labour market. Despite this, as reported by the OECD (2022), self-employment rates among migrants often surpass those of natives, highlighting a remarkable entrepreneurial spirit. This phenomenon persists despite significant obstacles that migrant entrepreneurs face, including limited access to finance, unfamiliarity with local business regulations, and insufficient social and professional networks (European Migration Network, 2019).

The entrepreneur's personal characteristics profoundly influence the path of entrepreneurship. The psychology of the entrepreneur plays a significant role in new venture creation, as explored by Baum et al. (2007). Entrepreneurial motivation is considered a key element in the entrepreneurial process, a topic thoroughly examined by Shane and Venkataraman (2000). An entrepreneur's personality traits, including risk tolerance and need for achievement, have been shown to shape business decisions and strategies (Rauch/Frese 2007). Research has also indicated associations between the Big Five personality dimensions<sup>1</sup> and entrepreneurial status (Zhao/Seibert 2006). Beyond individual traits, the individual's social connections are crucial for identifying and capitalising on opportunities (Davidsson/Honig 2003). Personal traits, connections, and business ideas are all crucial aspects

---

<sup>1</sup> Through the Big Five personality theory, personality is understood in a five-dimensional model, namely extraversion, openness to experience, emotional stability, conscientiousness, and agreeableness.

to be investigated and deepened to understand and adequately orient the choices of the aspirant entrepreneur, especially for migrant entrepreneurs who, in some cases, may lack established networks in their new environment.

To delve into the deep personality characteristics of migrant entrepreneurs, to support the specification of the business idea and thus be able to support it best, a literature review is not sufficient, which tends to find generalist and stereotypical information, but an added personal investigation of past, present and future self is needed. To support the understanding of the individual, we rely on their self-representation, understood as a dynamic product, which is well represented by the conception of the individual proposed by Napolitani (1978). In his conceptualisation, the "self" is represented by the dynamic interconnection between prior cultural, knowledge and relational baggage of the individual (the *Idem*) and the predisposition to a transformative understanding of the world, freed from the constraints of established culture (the *Autos*) (Napolitani, 1978).

This perspective underscores the importance of the entrepreneur's cultural and social reference system, together with their transformative capacity in interpreting and re-elaborating phenomena and symbols.

Self-narration is considered a valuable means to reach such an in-depth understanding. Discourse analysis as a means to understand the entrepreneur's perspective is rooted in the principles of constructivist social theory (Parkinson/Howorth 2008; Moser et al. 2019). This approach considers language the primary medium for producing meaning and, thus, a tool for shaping and being shaped by reality (Jacob 2004; Parkinson/Howorth 2008). Unveiling discourse, its meanings and their mutual interconnections can open a deep understanding of the individual, with the most accurate portrayal of one's representation of reality (ibid).

The representation revealed can be an extremely useful "tool" for entrepreneurs. Firstly, the restitution of these connections can increase awareness of the narrative they create about themselves and their business idea. Secondly, it can act as a medium for consolidating,

adapting or overcoming certain beliefs and representations of their business idea. Ultimately, this renovated understanding of one's own business idea groundwork can inform the design of customized entrepreneurial support models based on the entrepreneurs' unique and characteristic elements and the context in which it is embedded, improving personal data-driven strategies.

## 2. Aim of the research

The present research intends to illustrate the process by which meaning extraction and preliminary understanding can be formulated from a qualitative co-occurrence analysis on migrant entrepreneurs' self-narratives. For this aim, we will investigate how the subject entrepreneur interprets and represents their idea of “personal project” (or business idea) within their own discourse and, in doing so, what associations with other narrated elements of their own experience and environment it displays. Through the linguistic analysis of co-occurrences, the study aims at highlighting which concepts are closest<sup>2</sup> (i.e., those that are most co-occurring along the text, that is, those that are most likely to be found close and connected) but also those that are less close therefore to be considered more disconnected, less interdependent.

The scope of this procedure is to highlight unconscious and covered associations between ideas that, once unveiled, can provide to the speaker a deeper understanding of how their business idea is framed in respect to other concepts (e.g. real needs, personal traits, past experiences or desires) and to extract meaning from it for a more informed entrepreneurial support.

---

<sup>2</sup> As for the formal representation of such closeness, see pag. 41.

### 3. Methodology

The present study collected data through semi-structured interviews to stimulate spontaneous responses and minimise conditioning on discourse (Hollway/Jefferson 2000). The interviews were conducted among participants in a business pre-incubation programme for individuals with migrant backgrounds in Italy. Two interviews were selected as case studies for this research, conducted with Ada and Sarah, two women, respectively, from Albania and Liberia, aged 36 and 30 and representing different migrant generations (first and second). The criterion for selecting these interviews was based on two aspects. Firstly, the aim was to seek consistency in terms of gender (female), stage of business development (germinal stage) and level of Italian (native speaker); secondly, considering the pages at our disposal, we chose to deal with only two interviews.

The interviews were transcribed into textual form to allow the extraction of deeper meanings of self-narrative. The considered methods aim at automatically locating and extracting concepts within texts, a procedure which leads to Automatic Knowledge Representation (AKR). Being a procedure pertaining to discourse analysis and NLP, it requires the creation of specific lingware, as for instance formal local grammars in the form of Finite-State Transducers (FSTs)<sup>3</sup>. A second step consists in studying the word co-occurrences inside the texts analysed, in order to unveil and define the semantic relationships between different elements (such as words, phrases, or other linguistic elements), in order to represent the knowledge such texts include. This method is particularly relevant: it can enable the understanding on how certain narratives are constructed and reinforced over time, providing valuable insights about semantically-relevant word proximity, and also about the distance between different elements and concepts within discourses. To implement such a process the present study relies on the

---

<sup>3</sup> For further information on Finite-State Transducers, see Schutzenberger M. P., (1961) A remark on finite transducers. and Perrin, D. (1990). Finite automata. Handbook of Theoretical Computer Science

use of NooJ NLP Environment (Silberztein et al. 2016), a natural language processor structurally capable of parsing through different levels of linguistic phenomena (Ledin/Machin 2020). Following the procedure proposed by Olivieri et al. (2023), the first stage involved extracting and selecting meaningful tokens and syntagmas, which were subsequently grouped into concepts. In support of this conceptualisation, an open coding phase was conducted (Bryman 2012) independently for each interview, which allowed for the emergence of ideas and concepts that were specific to each speaker and not necessarily referable to pre-existing concepts (Olivieri et al. 2023). Through this process, for each interview, unique concepts emerge, although common and recurring themes are identifiable. These concepts and related tokens, within a constructionist perspective (Bączkowska 2010), are considered “construction grammars” which “themselves carry meaning, independently of the words in the sentence” (Goldberg 1995: 1): this imposes a careful and thorough qualitative understanding of the discourse, and results in a strong anchoring of the meaning of each concept to the context of each interview.

This understanding was supported by reference to Systemic Functional Linguistics (SFL) principles. Certain concepts are reflected by the *Three Metafunctions of Language* (Halliday/Hasan 1985; Machin 2016): *Ideational Metafunction*, which focuses on how language represents reality, allows the analysis of how entrepreneurs conceptualise their environment and personal experience understanding (e.g., “needs”, “personal experience”, “desires” and “support networks” concepts...); the *Interpersonal Metafunction*, which enlightens the type of relation established during the conversation, thus reflecting their identity and self-positioning in a certain environment (e.g., “the self” grammars; reporting linguistic structures that reflect one's self-perception and representation); the *Textual Metafunction* that shows the formal structure used to express and contain the previous two metafunctions, helping understanding how the entrepreneurs structure their narratives and linking ideas together (e.g., use of passive form, type of verbs employed, and so on.).

Each concept is then syntactically represented by a formal and local FST grammar (Silberztein, 2016) of the NooJ software. The formal and semantic grammars are displayed graphically in NooJ, based

on the standard score, allowing the creation of occurrence diagrams that show the frequency of the regular expressions comprising each concept along the text. A co-occurrence analysis is then conducted to compare the distribution of multiple concepts in each text. Through this software, two concepts are considered co-occurring if, within a portion of text equivalent to 1/20, they occur simultaneously to a significant and relevant degree. Through graphs in standard score, for the co-occurrence of the concepts in the text, it is possible to check the points of greater or lesser contact as well as intersection and/or divergence between them.

Referring to interpretive patterns in (Olivieri et al. 2023), such as “alternating peaks”, “concatenations”, “low frequencies”, and “overlapping words”, interrelationships between pairs of grammars within text segments are highlighted from the co-occurrence patterns. This approach allows us to analyse the conceptual closeness or distance between the topics proposed by the speaker, where closeness means a frequent and distributed co-occurrence between two concepts and distance is a low co-occurrence.

For each interview, a co-occurrence analysis starts with the grammar of “Personal Project” - the hat of all regular expressions related to the interviewees' business idea - with the other emerging concepts. That “Personal Project” is a grammar present in both interviews and is put in relation to other concepts emerging from the text (*community, challenges, needs, and goals* are some of them as an example). The goal of the analysis is to identify the concepts that most co-occur with those of “Personal Project,” as well as those that exhibit less conceptual proximity, thus not occurring discursively together.

Concepts with high co-occurrence to “Personal Project” reflect consistent co-existence within the discourse, highlighting how they accompany and connect to the entrepreneurial project concept. This strong connection provides the researcher with significant information about how the speaker interprets and represents their business idea, revealing which other concepts are integrated into the discourse in relation to the personal project.

Concepts that co-occur less frequently with “Personal Project” can still provide relevant information. As Burt (1992) described in his “theory of structural holes,” such concepts represent opportunities because they separate non-redundant sources of information that are

additive rather than overlapping. In this sense, weaknesses or conceptual gaps detected in the analysis may indicate potential spaces for development, improvement or investigation of the entrepreneurial project.

The chosen approach is based on the desire to map the speaker's meaning-making process. This will identify the meanings the speaker attaches to discourse concepts and understand how these relate to the intention to establish an enterprise. Such an analysis allows us to investigate how the speaker represents their business idea in relation to other elements of their experience as a basis for future targeted support.

### 3. The interviewees

*Ada:* An Albanian-Italian woman who immigrated to Italy at the age of ten with her family. As a young adult, she became involved in promoting Albanian culture in Italy, participating in and organising cultural events. In her migration path, she experienced the challenges of being an Albanian immigrant in Italy, particularly the issues of identification and cultural preservation. Her business idea is to commercialise traditional products, such as typical Albanian handicrafts, food, and products that carry a solid cultural symbolic value, to spread the culture of Albania all over Europe. The business has not started yet, but she is developing the idea and creating many significant relationships needed to start it up.

*Sarah:* An African-Italian second-generation young woman. She has experienced being “the only Black in the room” several times in her life, which has reinforced her desire to create more inclusive spaces. Still rather undecided about the final model of her business idea, Sarah wishes to create a space of inclusion, capable of enhancing the talents of migrant people and promoting their self-affirmation. She cultivates a passion for the worlds of fashion, literature and music, which inspire her business idea. Sarah’s background includes public speaking, coaching, and being a testimonial for fashion and social

projects. Her experiences have led her to develop a strong sense of mission in helping others discover their potential and overcome societal barriers.

#### 4. Analysis

For the purpose of analysis and subsequent reflections induced by each of the interviews, two co-occurrence studies will be reported: one concerning two close concepts (CASE 1) and the other involving two distant concepts (CASE 2).

##### *4.1. Ada*

###### *Case 1:*

Within Ada's interview, the first co-occurrence analysis concerns the semantic grammars of "Personal Project" (PP) (Fig. 1) and "Personal Goals" (PO). The concept of PP contains all the expressions that relate to Ada's business idea, thus elements referring to the target group, product type, and business structure, just to mention a few. On the other hand, the grammar of PO refers to the desires and goals that Ada aimed and still aims to achieve. The first observation about the trend of the two graphs (Fig. 2) concerns the presence of expressions related to the concept of "Personal Goals" in the first half of the text (particularly highlighted by the peaks in parts 5 and 10). In this first part, the two semantic grammars do not occur. This underscores how some of Ada's goals do not relate to her entrepreneurial project. These include, for example, "*And also I wanted to **return to Albania** (PO), not as my parents' daughter. But as someone else. As **an actor of development** (PO)*"; "*we had many meetings to try to **understand the needs**", "I wrote a biography that was a storytelling, with the help of colleagues, that also could **show the needs** (PO) from which this handicraft*

started”. However, these are just a few expressions. In concomitance with the introduction of the discourse about her business idea, Ada begins to share significantly more about her goals, which are constantly interconnected with her entrepreneurial project.

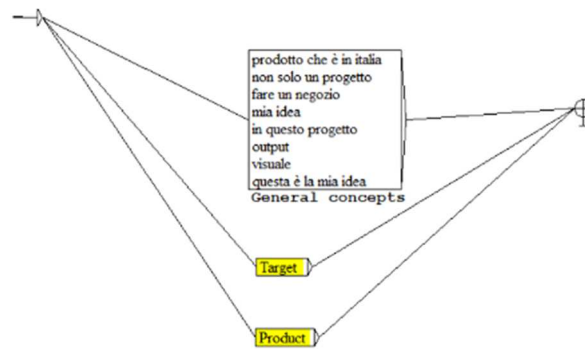


Figure 1. Ada “Personal Project” grammar.

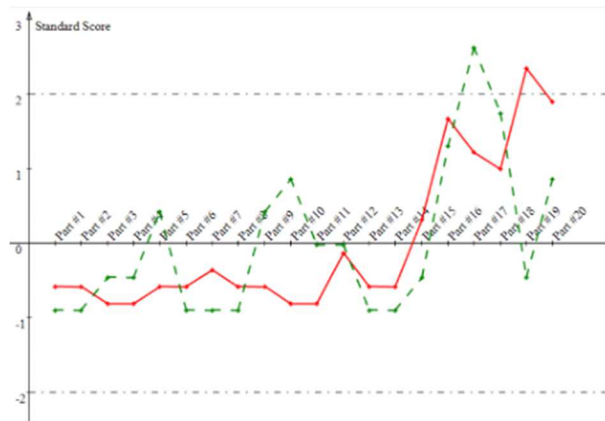
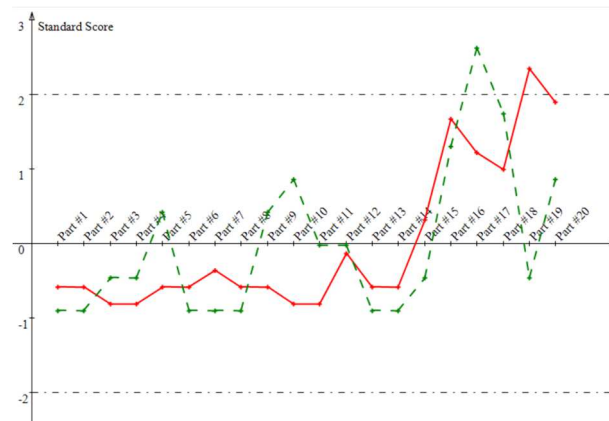


Figure 2. Co-occurrence diagram: Interviewee N.1 ‘Personal Project’ grammar (continuous line) [Fq=72] and ‘Personal Objectives’ grammar (dashed line) [Fq=41].



*Ada: The idea is to make it a **business** (PP) and to do these operations **on a regular basis** (PP), without feeling guilty, because maybe it's a person I love and I ask him for money, etc. And it would all be much **more systematic** (PP). So, 2020 made me wake up, made me realise that the **second generations** (PP), actually, **the real second generations** (PP) today of Albania don't know a damn thing! They have the fetish of Albania; however, they also **have to know** (PO) in order to speak in some way and they **have to be empowered** (PO) regarding that... and I also strongly believe in **accountability, even personal, political** (PO). I mean... as a **planned strategy** (PO, PP). These **second-generation** (PP) dancing, making videos on Tiktok... I am Albanian and proud to be Albanian, they even make songs...however, they should **be given a tool** (PO) to be useful. Another thing I discovered during the work with the incubator was that I don't want to drop the **social value of the project** (PO, PP)*

Quote 1. Interviewee N.1, Part 16: 'Personal Project' grammar (PP) "Personal Objectives" grammar (PO)

### Case 2:

The second co-occurrence analysis relates to the semantic grammars of "Personal Project" and "Past Experiences" (PE). As can be seen from the graph (Fig. 3), the grammar of "Personal Project" does not have a relevant presence in the first part of the text, while it is more crucial from part 14 onward. Effectively, before those passages of the interview, Ada makes very few explicit references to her own business

activity. In this case, for the grammar of “Personal Project”, the achievement of “Low standard score but high frequencies” (Olivieri et al., 2023) is not found either, since the standard score is too low. Indeed, in the first part of the text, Ada prefers to narrate her own story and her work and life experiences. The grammar of “Past Experiences” is very much present in the first two-thirds of the text, in alternating stages, until it nearly disappears starting from Part 15. The peaks in Part 4, 10 and 13 correspond with detailed accounts of certain periods of her life, where she talks about past projects she had to coordinate (“worked,” “returned to Albania,” “managed projects,”...) the challenges she faced (“the needs”, “criticisms”,...) and the skills she learned (“branding,” “training,” “promoters,” “campaigns,” “ethno-marketing”) interposed with specific reflections or digressions. The only significant point of “proximity” (Quote 2) that is observed between these two semantic grammars is the closeness between the tokens “criticisms” and “needs” (in the grammar “Past Experience”- also belonging to the grammar “Challenges”) and that of “*in this project*” (in the grammar of “Personal Project”) which opens on the predominant concept of her own business idea. This *alternating peaks* structure (Olivieri et al., 2023) seems to represent Ada's logical connection between her past and her own business idea, putting in causal sequence an element she experienced in her past with the reasons that move her own project. It should be specified that the slow transition that then connected the two discourses was stimulated by a question from the interviewer that coincides precisely with part 15.

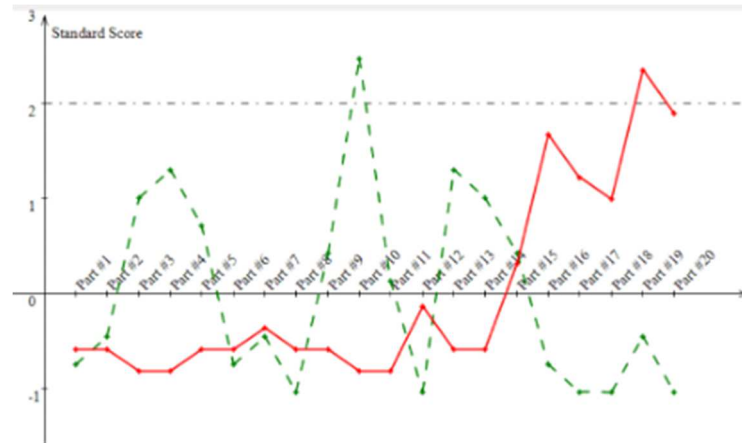


Figure 3. Co-occurrence visualisation between “Personal Project” (straight line) [Fq=72] and “Past Experience” (dashed line) [Fq=74].

**Ada:** Yes, absolutely, otherwise I would not have told all these things. So, to me, all **the disappointments (PE)** I had during **this project design (PE)** led me to the idea that... they made me feel a little bit guilty about it, this **instrumentalisation of the artisans (PE)** who then don't gain anything from it [in terms of money]. However, their experience is valid, [...] That is, you have to listen to it. And so, on the one hand, all the things, **criticisms and needs (PE)** - let's say a map of the needs that they expressed - is in my opinion important, **in this project (PP)**.

Quote 2. Interviewee N.1, Part 15: ‘Personal Project’ grammar (PP) ‘past working experience’ grammar (PE)

## 4.2. Sarah

### Case 1:

In Sarah's interview, the case of co-occurrence with the highest distribution in the text concerns the grammars of “Personal Project” (PP) and “Self-Narration” (SN), which could be summarised as all the expressions used by Sarah to refer to the self. Among them we find, for example, “*I realised this,*” “*I love challenges of course*”, “*because I feel in the middle (SN), between these two realities, one half Italian (SN),*

*and half African (SN).*” A solid co-occurrence between the two semantic grammars is evident from the graph (Fig 5). Even more in-depth parsing of the text brings out a constant reference to her own characteristics and a strong self-awareness, that goes along with the reflections on her own business idea. This creates a strong connection between her own individual characteristics and those that her future business idea should have.

*Sarah: Yes, Actually I've been thinking a lot about this for the simple fact that I've worked with both **trafficked women** (PP) and I think they are the most difficult target group. They are the principal challen...I mean eh **I don't think I can** (SN) work with them, although... I mean no, wait! What am I saying?! **I love challenges** (SN) of course, the bigger they are the better I am.*

*In the sense, I'm happier [with bigger challenges] and so **trafficked women** (PP) are actually **of interest to me** (SN). But in order to work with them I know **I need a team** (PP) that is present, strong, equipped. (Part 7).*

*Sarah: But all this had started, I'll tell you the truth, from the simple idea of making this **fashion collection** (PP) which is a collection that I think, compared to all the other clothing lines seen so far, where they use WAX material, which is so, so...*

*Cecilia: very, stereotypical. you say?*

*Sarah: Stereotyped. I wanted to create something that kind of **looks like me** (SN), which is **always Africa** (SN), because it is always Africa, however it's something **more artistic** (SN), so **each piece** (SN) has to be a work of art, okay? (Part 16)*

Quote 3. Interviewee N.2, Part 7 and 16: “Personal Project” grammar (PP) [Fq=61] and “Self-Narration” grammar (SN).



Figure 4. Sarah's "Personal Project" Grammar.

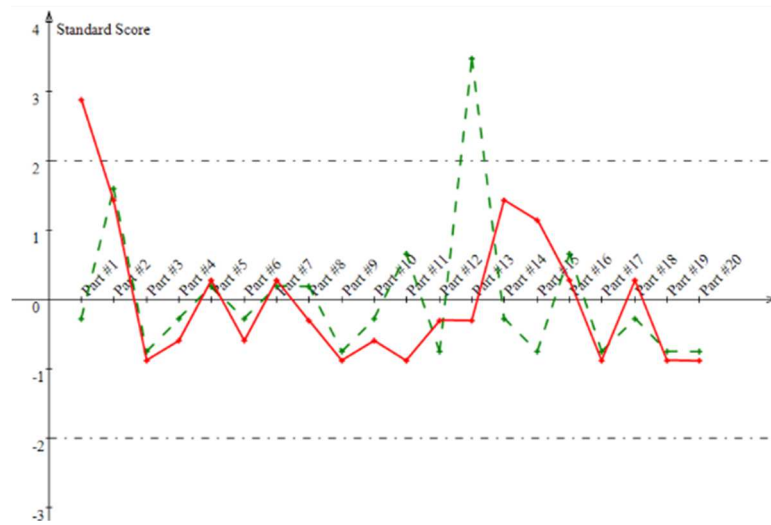
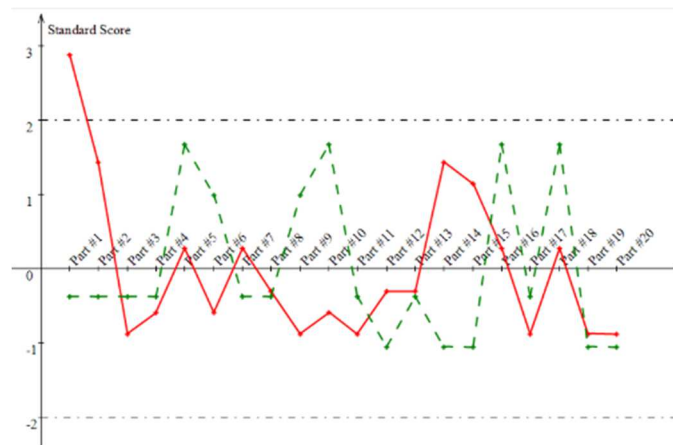


Figure 5. Co-occurrence visualisation between "Personal Project" (straight line) [Fq=61] and "Self-Narration" (dashed line) [Fq=32].

### Case 2:

The second co-occurrence study to be considered in Sarah's interview concerns the distribution of the concepts of "Personal Project" and "Personal Objectives" (PO). The near alternating pattern between the two semantic grammars is of particular relevance, as it could reflect either a sequence of concatenated concepts, typical of *alternating peaks* structures (Olivieri et al., 2023) or a semantic distance between the two concepts. From a closer analysis of the text, it is recognised to be the second case. Sarah in fact proposes along the text many references to her own personal and life goals, but without putting them in direct correlation with the idea of enterprise she is talking about, hence a lack of co-occurrence between the two semantic grammars. If we look at two points of low but existing co-occurrence, we notice how poorly defined the idea of enterprise she has is, in contrast to the vividness of her own almost "moral" goals.



**Sarah:** So honestly, I don't know *what idea of project* (PP) I have, but one thing is sure: *lighting that flame* (PO) in other people is *my mission* (PO). Then I wouldn't know how to answer the question about *what idea of enterprise* (PP) I have. I mean, because in my opinion *it could have another form*. (Part 10)

**Sarah:** Yes, yes, I would see *this atelier* (PP) more as *a house* (PP) than an atelier.

*And yes, I have to say that another aspect is the following: I experience this **as a life mission** (PO), I think that, if I went through all that, and still internationally they saw me **as an activist** (SN) or a **person who made it** (SN) despite my background... In my opinion, I mean I feel that **I have to dedicate** (PO) even more of my life and **my existence** (PO) to others.*

*In which sense? Because **I enjoy this privilege today** (SN) I have a privilege, because I only had to enter the International Book Show space to understand that. Being the **only black writer** (SN), I mean, in an international show like that, makes you realise that **you are the privilege** (SN). And that's a privilege that used to **make me die inside** (SN).*  
(Part 13)

Quote 4. Interviewee N.2, Part 10 and 13: 'Personal Project' grammar (PP), 'Self-narration' grammar (SN), and "Personal Objective" grammar (PO).

Figure 6. Co-occurrence visualisation between "Personal Project" (straight line)[Fq=61] and "Personal Objectives" (dashed line) [Fq=31].

## 5. Results

Based on the above graphs and the information obtained, it is possible to proceed toward considerations of an "action" nature. This means extracting actionable insights (Kemmis et al. 2014) and orienting value for the speaker in their journey of entrepreneurial understanding and growth.

Ada's interview unfolds in two distinct moments. In the first part, past life experiences are recounted in great detail, in which Ada devotes a great deal of insight to her migration experience in Italy and her career paths, accompanied by their respective difficulties (Fig 3, until Part 15). Therefore, this narrative is interrupted when she moves on to the narrative of her own business idea and related goals. Almost no significant co-occurrence is observed between the concepts of "Personal Project" and "Past Experiences", an aspect that seems to anticipate a general narrative disconnection between one's past and the entrepreneurial activity she intends to start. At the same time, an expression afferent to the grammar of "Personal Project" (the target group "second-generation youth") is mentioned three times throughout

the narrative of her past. This is a small number compared to the size and frequency of the grammar within the text. However, it reveals how reflections and contacts with this social group have been traced back to years before the interview or the conception of the business idea. It is, therefore, possible to suspect a growth of awareness that has matured over the years, which, however, is scarcely marked in the text. The first insight that can be gained from this analysis of (missed) co-occurrence concerns the invitation to carefully examine past experiences to define and consolidate Ada's business idea. Indeed, the strong emphasis on her past suggests its great relevance to Ada, which is why goals, awareness, and resources for defining the business idea should be further investigated in this "part" of her story. A second aspect to consider is the close co-occurrence observed between Ada's personal objectives and her business idea in Fig. 2, starting in Part 15. From that point on, within the discourse, personal goals and projects are very solidly matched, showing a certain coincidence, unlike in Parts 3 and 10, where personal goals are clearly unrelated to her project. This strong co-occurrence highlights which goals Ada surely aspires to achieve through her entrepreneurial project, setting some precise benchmarks on the trajectory of her business and what elements she will have to fulfil.

Sarah's journey encompasses a wide range of personal and professional experiences, all highly consistent with her worldview and self-attributed sense of responsibility in fostering new stimulation and resources for young Afro-descendants. Moreover, like these experiences, the desire to start a business idea is constituted in line with these values. However, while analysing her narrative in the text, it was observed that when talking about her own life and values aspirations, Sarah mentions elements related to her business idea infrequently (Fig. 6). This aspect, reinforced by the weak and vague expressions referring to her own business (e.g., Quote 4, Part 10), seems to enhance Sarah's determination to pursue those goals regardless of what form or characteristics her business idea will take. Thus, we could say that the enterprise for Sarah reflects a vocation rather than a profession in that it tends to one's aspirations, which are understood as a life purpose. This analysis thus allows for the assumption that Sarah is open to various

business opportunities and may even tend toward entrepreneurial seriality, as one's idea is not necessarily rigid but is open to modifications, even substantial ones, as long as the social purpose remains at the centre. The beneficiaries/clients of the project, like the mission, will never vary much: the realisation of her business, whatever it will be, must be devoted to the involvement of marginalised subjectivities. The “rigidity” of the characteristics of the target audience, along with the flexibility of the business idea, are features found in several social enterprises (Borzaga 2009): Karima's seems to be precisely compatible with this type of business that aims at generating a positive social impact, rather than to purely economic growth. At the same time, an essential aspect to consider in future entrepreneurial support is the close correlation between the business idea and one's self-narrative (Fig. 5). The strongly co-occurring trend of the two concepts within the entire text tells of the central role of Sarah's experiences, personality, and beliefs in shaping the entrepreneurial form. For this reason, specific support should be directed toward identifying the elements of the self that should be foundational building blocks for defining her idea.

## 5. Discussion: Evaluation of the Research

Two main limitations can be identified within the present study. The first limitation is technical and involves the level of linguistic disambiguation required in constructing semantic and syntactic grammars. The study utilises extended tokens to capture specific linguistic points within the text, resulting in a highly granular and qualitatively rich data set. However, this method, necessitated by the high level of disambiguation required, limits the semantic grammars' adaptability to other texts. Consequently, the semantic grammars developed in this study are customised for a specific context and are tailored to each interview, making it challenging to apply them more broadly across other corpora without adjustments. This methodological

choice prioritises individual-focused insights over identifying broader patterns, a direction for future testing and refinement. The second one concerns the analysis process. In both cases, the proposed considerations are gathered based on the relationship between just two pairs of concepts. The purpose of the present work was to validate a procedure already formalised in Olivieri et al. (2023) in investigating possible insights and awareness about the individual and one's conception of business from the (failed) co-occurrence between two concepts: in future action studies and programmes, adequate and comprehensive support must be based on an inclusive analysis of all relationships and interdependencies among the concepts that emerge. This is the only way to obtain a representative picture of the entrepreneur's representation of their own experience and idea of business on which to intervene for support.

Then some methodological considerations should be made. The level of deepness and granularity of the data accessible through this process suggests that NooJ is a valuable tool for hosting such qualitative analysis, by providing the appropriate formal environment for formalised and scientifically supported results (Silberztein 2016). The open coding procedure followed for each of the interviews and the systematic review (Gopalakrishnan/Ganeshkumar 2013) is rigorous and accurate, and aims to represent an interpretation that is as accurate as possible with the system of meanings attributed by the speaker. However, there is an awareness that absolute adjacency between the interviewer's interpretation and the meaning attributed by the speaker to a certain word or concept cannot be proven: belonging to different communities and considering language a socially constructed process, the meanings attributed to the same expression might vary (Chen 2015), possible alterations in the speaker's expression due to social conditioning (Grimm 2010), or linguistic differences reflected in a misleading syntax could occur. It is for this reason that a process of restitution and validation with the speaker is indispensable. This process reports multiple methodological benefits. First, through the restitution of research results by the speaker, which can result in validation, rejection or rediscussion, a qualitatively more refined conceptualisation of linguistic data is achieved. Second, the involvement of the respondent within the process of data collection and validation opens up a practice peculiar to Participatory Action Research

(PAR), a methodological approach that will also be implemented in the proceedings of the present study. PAR practices allow active participation by the respondent who, from being the “object” of the research, becomes an active “subject,” re-acquiring agency within the research scheme (Kemmis et al. 2014; Lenette 2022). Finally, a final benefit is observed when considering the ultimate goal of the present study. The understandings obtained from the analysis play an essential role within the entrepreneurship coaching pathways in which respondents participate. As specified earlier, discourse analysis offers important and sometimes hidden - complex information about the meanings the speaker attributes to their experiences. Receiving such feedback can mean dealing with unconscious processes of that representation for the respondents, over which they regain control through validation, rejection, or questioning. Through this process - which will have to be repeated on several occasions during the support process - it's possible to proceed toward the gestalt evolution of individual awareness (Çeliköz et al. 2019), functional to shape their business idea as authentically as possible, supported by a thorough understanding of their individual experience and the role it plays in the development and consolidation of the business idea.

Since this is potentially sensitive data collected through interviews, strict ethical standards were upheld. The research purpose was clearly explained to participants: informed consent was then obtained, particularly regarding the use of a voice recorder. To protect confidentiality, participants' identities were anonymised during data analysis. Additionally, leading questions were avoided to prevent bias and ensure the integrity of the information collected.

## 7. Conclusion

This study enhances the value of leveraging discourse analysis through tools like NooJ to uncover the nuanced self-representations and aspirations of migrant entrepreneurs. By exploring the co-occurrence patterns within personal narratives, it is possible to gain critical insights

into how these individuals construct their entrepreneurial identities and envision their personal projects. It emerges how entrepreneurship and incubation programmes can benefit from such a process in order to intervene more accurately and through a more grounded approach in the individual business development support. For instance, in Ada's case, the presence of weak but existing co-occurrence between the narrative about her "past experience" and her business project just introduces the relevance she attributes to her past experiences for her business idea's choice; this highlights the need to further explore her past, to solidify her entrepreneurial origins, resources and vision. In contrast, Sarah's narrative reflects a flexible and mission-driven approach, emphasising her commitment to social impact over rigid business models. These differences suggest that tailored support systems must account for individual trajectories, emphasising personal histories, values, and aspirations rather than favouring predefined support schemes.

The importance of effective, context-sensitive and individual-centred entrepreneurship support strategies is one of the main objectives of this study. The integration of PAR will be imperative in order to maximise these cornerstones. Indeed, PAR encourages the involvement of migrant entrepreneurs in the analysis and validation of their narratives, ensuring that the insights derived are not just accurate and grounded but also empowering. Indeed, this process also fosters self-awareness and agency, enabling participants to critically reflect on their entrepreneurial backgrounds, resources and aspirations and align them with actionable goals.

As mentioned above, the method proposed in this study takes into account one last crucial element: the business environment. The embeddedness with the environment and the context in which the entrepreneur intends to enter constitutes an element of strong influence on the entrepreneur's choices and orientations (McKeever et al. 2014): mapping and understanding the entrepreneur's representation of their own environment can foster greater consistency of intervention, which takes into account and is responsive to environmental changes. The iterative nature of PAR, integrated with the method here exposed, allows for continuous feedback and adaptation, creating a dynamic support framework that evolves alongside the entrepreneurs' needs, experiences and environment.

The implementation of this discourse-grounded support system, integrated by participatory frameworks, will foster authentic and more sustainable business pathways that align with their entrepreneurs' unique identities and aspirations and with the context they are situated in. Through such methodologies, it is possible to create entrepreneurial ecosystems that are not only inclusive but also transformative, addressing complex environments while enhancing individual potential.

## References

- Bączkowska, Anna 2010. Constructivism and constructionism in language and cognition. In Puppel, Stanisław / Bogusławska-Tafelska, Marta (eds) *New Pathways in Linguistics*. Olsztyn: University of Warmia and Mazury, 9-20.
- Baum, J. Robert / Frese, Michael / Baron, Robert A., & Katz, J. A. (2007). Entrepreneurship as an Area of Psychology Study: An Introduction. In Baum, J. Robert / Frese, Michael / Baron, Robert A. (eds) *The psychology of entrepreneurship*. New York: Lawrence Erlbaum Associates Publishers, 1-18.
- Borzaga, Carlo 2009. L'impresa sociale. In Bruni, Luigino / Zamagni, Stefano (eds) *Dizionario di Economia Civile*. Roma: Città Nuova, 516-26.
- Bryman, Alan 2012. *Social Research Methods*. 4th. New York: Oxford University Press.
- Çeliköz, Nadir / Erişen, Yavuz / Şahin, Mehmet 2019. Cognitive learning theories with emphasis on latent learning, gestalt and information processing theories. *Journal of Educational and Instructional Studies in the World* 9/3, 18-33.
- Chen, Bo 2015. Social constructivism of language and meaning. *Croatian Journal of Philosophy* 15/43, 87-113.
- Davidsson, Per / Honig, Benson 2003. The role of social and human capital among nascent entrepreneurs. *Journal of Business Venturing* 18/ 3, 301-331,

- European Migration Network 2019. Migratory pathways for start-ups and innovative entrepreneurs in the EU. < [https://home-affairs.ec.europa.eu/system/files/2022-09/migratory-pathways-for-start-ups-and-innovative-entrepreneurs-in-the-eu\\_2019\\_BG.pdf](https://home-affairs.ec.europa.eu/system/files/2022-09/migratory-pathways-for-start-ups-and-innovative-entrepreneurs-in-the-eu_2019_BG.pdf)>
- Goldberg, Adele 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.
- Grimm, Pamela 2010. Social Desirability Bias. In Sheth, Jagdish / Malhotra, Naresh (eds) *Wiley International Encyclopedia of Marketing*. Hoboken, NJ: John Wiley & Sons.
- Halliday, Michael A.K. / Hasan, Ruqaiya 1985. *Language, Context, and Text*. Oxford: Oxford University Press.
- Hollway, Wendy / Jefferson, Tony 2000. *Doing qualitative research differently: Free association, narrative and the interview method*. London: Sage.
- Kemmis, Stephen / McTaggart, Robin / Nixon, Rhonda 2014. *The Action Research Planner: Doing Critical Participatory Action Research*. Singapore: Springer.
- Ledin, Per / Machin, David 2020. *Introduction to multimodal analysis*. London: Bloomsbury Publishing.
- Lenette, Caroline 2022. *Participatory Action Research: Ethics and Decolonization*. Oxford Research Encyclopedia of Methods.
- Machin, David 2016. The Need for a Social and Affordance-driven Multimodal Critical Discourse Studies. *Discourse & Society* 27/3, 322-334.
- McKeever, Edward / Anderson, Alistair R. / Jack, Sarah 2014. Social Embeddedness in Entrepreneurship Research: The Importance of Context and Community. In Chell, Elisabeth / Karatas-Ozkan, Mine (eds) *Handbook of research on small business and entrepreneurship*. London: Edward Elgar Publishing, 222-236.
- Moser, Susanne / Meerow, Sara / Arnott, James / Jack-Scott, Emily 2019. The Turbulent World of Resilience: Interpretations and Themes for Transdisciplinary Dialogue. *Climatic Change* 153/1, 21-40.
- Napolitani, Diego 1978. *Individualità e gruppaltà*. Torino: Bollati Boringhieri.

- OECD 2022. *International Migration Outlook 2022*. Paris: OECD Publishing.
- Olivieri, Cecilia / L. Laquidara, Lorenzo M. / Semlali, Agathe 2024. Semantic Analysis of Migrants' Self-Entrepreneurship Ecosystem Narrative. In Bartulović, Anita / Mijić, Linda / Silberztein, Max (eds) *Formalizing Natural Languages: Applications to Natural Language Processing and Digital Humanities*. NooJ 2023. *Communications in Computer and Information Science* 1816, 176-187.
- Parkinson, Caroline / Howorth, Carole 2008. The language of social entrepreneurs. *Entrepreneurship & Regional Development* 20/3, 285-309.
- Perrin, Dominique. 1990. Finite automata. *Handbook of Theoretical Computer Science*, pages 2-56.
- Rauch, Andreas / Frese, Michael 2007. Let's put the person back into entrepreneurship research: A meta-analysis on the relationship between business owners' personality traits, business creation, and success. *European Journal of Work and Organizational Psychology* 16/4, 353-385.
- Schutzenberger, Marcel Paul. 1961. A remark on finite transducers. *Information and Control*, 4:185-187
- Shane, Scott / Venkataraman, Sankaran 2000. The Promise of Entrepreneurship as a Field of Research. *Academy of Management Review*, 25/1, 217-226.
- Silberztein, Max 2016. *Formalising natural languages: The NooJ approach*. Hoboken: John Wiley & Sons (2016).
- Zhao, Hao / Seibert, Scott E. 2006. The Big Five personality dimensions and entrepreneurial status: A meta-analytical review. *Journal of Applied Psychology* 91/2, 259-271.
- Zweig, David / Webster, Jane 2004. What are we measuring? An Examination of the Relationships between the Big-five Personality Traits, Goal Orientation, and Performance Intentions. *Personality and Individual Differences* 36/7, 1693-1708.



AGATHE SEMLALI

## Linguistic Analysis for Project Evaluation through NooJ: the EU WeMED\_NaTour Pilot Trip Assessment

### Abstract

The EU WeMED\_NaTOUR project aims to develop transnational eco-smart tourism packages targeting coastal and natural marine destinations in the West Mediterranean. The goal is to enhance the attractiveness of less-known destinations, with a special focus on younger generations. The project wants to enhance stakeholders' capacity building by integrating sustainable tourism concepts into school trip curricula and developing comprehensive guides for sustainable practices. A pilot trip is an integral component of the project, involving practitioners in the tourism industry, beneficiaries from schools, and local SMEs. This research aims to measure the impact the eco-educational trips had on students. During the project, questionnaires are designed to collect meaningful insights from the students. The collected data from the answers enables the evaluation of the project implementation to further fine-tune the product features based on the beneficiaries' desiderata. Our analysis will be processed through NooJ software (Silberztein, 2016), in which NLP procedures will be performed: through word frequency, co-occurrences and semantic-syntactic analysis sentiments, attitudes and opinions will be unveiled. Such routines enable one to perform a Sentiment Analysis characterised by a deeper comprehension of the audience needs and perspectives. The outcome of this study will be used to develop future eco-tourism strategies.

Keywords: Eco-Smart Tourism, NLP, Project Evaluations, Questionnaire, NooJ

## 1. Introduction

The EU WeMED\_NaTOUR (WeMED for short) project, funded by the European Union, intends to focus on West Mediterranean natural marine sites and coastal destinations. WeMED aims to develop transnational eco-smart tourism packages mixing sustainable, responsible, and eco-tourism to increase the attractiveness of the destinations in the participant countries, highlighting their unique value to a broader audience of visitors, with specific reference to the “new generations”.

The project key goals include (a) enabling tourism SMEs to capitalise on the growing ecotourism market, (b) advancing innovation and digitalisation in the tourism industry to drive economic growth, (c) improving the business environment for SMEs, (d) diversifying tourism offerings, and (e) integrating eco-tourism concepts into school trip curricula (EU WeMED\_NaTOUR, n.d.). These primary goals were achieved through initiatives such as a capacity-building program, interactive workshops, educational materials for teachers, and eco-smart pilot trips, which provided students with immersive experiences in sustainable tourism. This flagship initiative comprises a consortium of partners from five countries, Italy, Belgium, Spain, Portugal, and Mauritania, coordinated by X23.

This article focuses on the pilot trips and their effects on their primary recipients, namely the students. Field trips are expected to foster knowledge gain and stimulate interest (Behrendt / Franklin 2014). The WeMED project organised four eco-smart pilot trips to immerse students in practical sustainable tourism experiences through ‘learning by visiting’ (European Climate, Infrastructure and Environment Executive Agency, n.d.; Falk et al. 2012). The trips were held in Italy, Spain, and Portugal in May 2024, while the fourth trip was held in Mauritania in October 2024. This research focuses on the trips to Spain, Portugal, and Italy, as the data from Mauritania was collected at a later stage. The destinations included in this paper were:

- Italy - Grado: Seventeen Spanish students (ages 14-15) participated in the Grado trip. The students explored sustainable tour-

ism through a city tour and activities on plastic waste at the airport. There was also a beach clean-up and a visit to a natural reserve, where students learned about coastal ecosystems and sustainable fishing.

- Spain - Palma de Mallorca: Twenty Italian students (ages 15-17) attended the Palma de Mallorca trip. The trip focused on wildlife conservation and discussions on protecting marine ecosystems. Students also learned about regional ecosystems through outdoor activities. Additionally, the trip included cultural exchanges with locals, promoting intercultural understanding and reflection on the trip-lessons.
- Portugal - Ria de Aveiro: Seventeen Portuguese students (ages 11-13) explored the Ria de Aveiro region. They also combined environmental and cultural education and took a solar-powered boat trip to learn about sustainable practices in Salicornia plantations and oyster farming.

The research provided insightful findings concerning the pilot trips involving high student engagement in sustainable tourism-related activities.

## 2. Purpose of the research

The overall WeMED research project aims to evaluate the level of implementation with which each stakeholder has achieved the objectives of the EU WeMED NaTour project. More specifically, this article assesses how eco-educational trips affected the students and were perceived and understood by them. We evaluate the effects of eco-smart educational trips by analysing specific keywords, phrases, and grammars and their interconnections related to environmental responsibility, cultural sensitivity, and responsible travel behaviors. We also assess the effectiveness of the trips in fostering sustainable tourism mindsets

among the next generation. These insights will provide valuable guidance for refining the WeMED NaTour project and maximising its impact on the regional tourism industry and environmental sustainability.

### 3. Research methodology

NooJ is an analytical software that enables deep written text analysis (Silberstein, 2016). It allows the researcher to identify keywords, concepts, recurrent words, and expressions and build grammars for Sentiment Analysis.

The research participants are purposely selected based on their participation in the pilot trips. The units of analysis of this research are the students' survey answers. To grasp their perspectives concerning the pilot trips, the surveys were completed shortly after them (Ballantyne / Packer, 2002). All participants were informed that their responses would be processed for research purposes. As all the students involved in the trips were minors, adequate information was shared with their parents to obtain their consent.

The WeMED project collected data through surveys, video, and audio recordings. For the sake of the pilot trip analysis here, only the survey answers were analysed. The survey used quantitative and qualitative methods, such as scalable questions and short written answers. Data collection through surveys is cost-effective and time-efficient (Bihu 2021). The surveys were tailored to the respondents' age and knowledge. The corresponding partners translated the surveys into each language to collect the answers in the respondents' native languages (Italian, Portuguese, Spanish). After collection, the data was translated into English for homogeneity and to ease the data analysis process for the researcher. The answers of the Italian students were not translated as members of the research team are native speakers. For translated data, a rigorous triangulation was performed with native speakers to ensure no meaning was distorted or mistranslated. The translators were all partners of the projects and hence had a comprehensive overview of the

project and extensive knowledge of sustainable tourism practices. Data is thus understood in its context.

The corpus of answers is divided by country. The surveys are first processed with NooJ by performing tokens extraction, in which all the tokens are retrieved, listing the words written in the answers of the surveys and indicating their frequency of occurrence. The researcher manually indexes the tokens to build positive and negative grammars inductively. In short, the tokens list is cleaned from stop words, and the remaining relevant words are selected based on their underlying emotional tone (positive or negative). This Sentiment Analysis determines how the students felt about the pilot trips. Assessing the current satisfaction of the pilot trips will contribute to fine-tuning the design of future sustainable school trips. Hence, based on the gathered data and the present methodology, we have primarily selected three NooJ analysis applications. The subsequent clarification delineates the categories of analyses that can be conducted.

- *Word frequency analysis.* Through token extraction, word frequency allows the researcher to identify and quantify the text's most frequently used words (George Washington University Libraries, n.d.). This mechanism helps gauge thematic emphasis and key concepts within the corpus by identifying topics that resonate most with stakeholders. It also previews the overall sentiment and helps build the grammars for the sentiment and co-occurrence analyses.
- *Sentiment Analysis.* It detects whether a piece of text is positive, negative, or neutral. Also called opinion mining, this technique is traditionally used for customer feedback and understanding their needs and opinions. This approach guides the researcher in detecting the respondents' emotions (e.g., surprised, shocked, happy, angry) and intentions (e.g., motivated, not interested). Sentiment analysis is performed by deep analysis of the text and syntactic recognition (e.g., negative forms—" [pronoun] do not [verb]...", for instance, "I do not like this") of positive, neutral, or negative linguistic forms through grammars construction and application (Monteleone et al. 2021).

- *Co-occurrence analysis.* It identifies grammars in the texts by grasping typical contextualisation, interpretation, and language usage (Olivieri et al. 2024). This research uses co-occurrence analysis to compare the positive and negative grammars and their dispersion throughout the survey answers.

#### 4. Students' evaluation and feedback of pilot trips

This section exposes the data and its interpretation for each country (Portugal, Spain, Italy) based on the surveys administered to the pilot trip students.

##### 4.1. Portuguese students

The fifteen survey respondents are 11-13-year-old students from Portugal traveling in Portugal.

##### 4.1.1. Key themes

The Words Frequency Analysis enables the identification of the most frequently used words in the questionnaires' transcripts. It identifies what topics resonate most with the students and previews the overall sentiment. It further helps to build the grammars to detect the positive and negative satisfaction trends.

A frequency-based analysis allows us to visualise the main themes discussed by the students during the surveys (see Table 1). For the Portuguese students, we can first remark that themes such as tourism and concerns for the planet are evoked. Additionally, components of the trips were significantly mentioned, such as the mode of transport, the boats, and the use of solar energy.

|           |        |         |           |             |          |
|-----------|--------|---------|-----------|-------------|----------|
| Tokens    | learnt | boat(s) | tourism   | environment | activity |
| Frequency | 20     | 19      | 11        | 11          | 10       |
| Tokens    | trip   | solar   | transport | planet      | protect  |
| Frequency | 10     | 7       | 6         | 6           | 6        |

Table 1. Most frequent relevant tokens.

#### *4.1.2. Sentiment analysis*

The grammars were constructed based on a thorough reading of the surveys and disambiguated to detect positive and negative. The words building the grammars were extracted from the answers and associated with positive or negative tones (neutral tone was excluded). This ensures that grammars directly reflect the positive and negative aspects of the trips. For example, the negative grammar contains words like 'boring', which indicates negative feelings about certain activities, or expressions like 'be longer', which show that some students wanted to stay longer and found the trip too short. In terms of disambiguation, for example, “least” is generally negatively connotated; however, in the answers, it was sometimes used in neutral sentences. Hence, the researcher disambiguated the token “least” into three different expressions where “least” was used negatively: “The least”[...], “My least”[...] and “found least”[...]. Thus, the grammar can detect whenever “least” expresses a negative feeling toward the trip. For instance, two different students expressed the following statements: “The least interesting activity was the BioRia” (Anonymous, Portugal Pilot Trip Survey, 2024), “The least interesting activity was the Bioria because the natural area is a bit degraded.” (Anonymous, Portugal Pilot Trip Survey, 2024). In addition to words and expressions, it is possible to find numbers (1 to 5), as evaluation scales were used in the surveys (1 stands for lower scores, such as 'Very dissatisfied' and 5 is the highest, such as 'Highly satisfied').

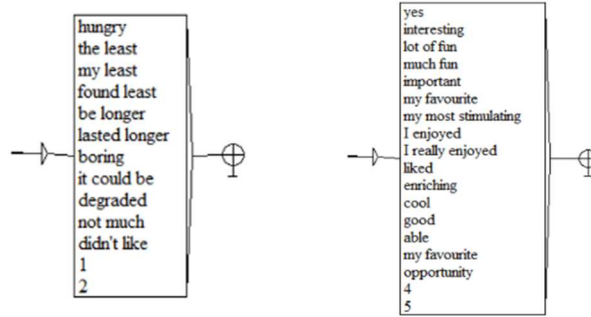


Figure 1. Negative and positive (from left to right) grammars for the Portuguese students' surveys.

The positive grammar also uses adjectives to define the quality of the trip (e.g., “interesting” and “stimulating”). This grammar enables the detection of highlights of the students' favorite activities, such as painting the boat, taking boats, skateboarding, and visiting the museum.

When the positive grammar was applied to the text, 123 tokens were located, meaning positive words and expressions were mentioned 123 times throughout the surveys. For the negative grammars, we located 18 tokens. To have more nuanced views when comparing both grammars, it is essential to note the ‘weight’ of positive or negative words (that is, the number of times they are mentioned throughout the text).

#### 4.1.3. Co-occurrences analysis

Co-occurrence analysis further deepens the sentiment analysis by comparing the positive and negative trends. In this graph comparison, we can see most of the peaks in the first half of the graph. It shows how, in the beginning, the students are expressing their opinions and satisfaction towards the trip, and by seeing peaks from both colors, we can deduce different and conflicted positions. Indeed, when asked about their overall experience, the kids explicitly described what positively struck or bothered them. Here, most negative comments are also related to elements of the organisational aspect (e.g., dissatisfaction towards the

guide, disturbance from their classmates at night, desire for more musical engagement, additional refreshments and treats), which are instead elements out of the control of the pilot trip design itself.

Regarding positive aspects, most activities were appreciated, especially outdoor ones. The absence of peaks between parts 9 and 14 (see Fig. 2) is related to the questions being oriented to the learning outcomes, hence having a more descriptive character concerning their new sustainability knowledge. Most students affirm gaining new sound understanding and that the trip will impact their ways of consuming and traveling: “I learned that I can contribute to reducing our ecological footprint by using environmentally friendly means of transport and playing with reused wooden toys.” (Anonymous, Portugal Pilot Trip Survey, 2024).

Overall, the graphs show that both the positive and negative grammars co-occur and peak around the same time (beginning and end). Indeed, when asked about feedback and improvement, the students gave nuanced and critical answers. However, positive feedback remains predominant, as we can see from the frequency of positive tokens (123) compared to negative ones (18). In brief, the Portuguese students felt positive and satisfied with the pilot trip.

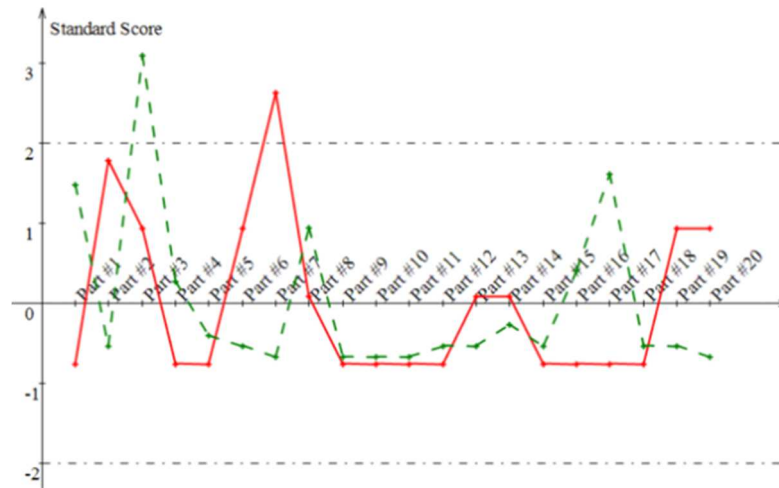


Figure 2. Graph comparison of the positive (full) and negative (dash) grammars across the students surveys.

#### 4.2. Spanish students

The fourteen survey respondents are 14-16-year-old students from Spain travelling to Italy.

##### 4.2.1. Key Themes

The following table illustrates some of the ten most mentioned tokens throughout the surveys of the Spanish students traveling to Grado, Italy:

|           |           |         |           |         |             |
|-----------|-----------|---------|-----------|---------|-------------|
| Tokens    | nature    | reserve | visit     | learned | environment |
| Frequency | 15        | 14      | 12        | 11      | 10          |
| Tokens    | different | rubbish | important | plastic | pollute     |
| Frequency | 10        | 8       | 8         | 7       | 6           |

Table 2. Most frequent relevant tokens.

The table suggests some relevant topics for the students, such as the environment and the nature they discovered in a new country and the impact of pollution (e.g., “plastic” and “pollute”). We can also grasp some vocabulary related to the concrete aspect of sustainability, the activities they experienced, and the lessons they learned.

##### 4.2.2. Sentiment analysis

The positive and negative grammars were constructed on the tokens listed in NooJ. A higher level of disambiguation was necessary in the Spanish case, as many words were mentioned in either positive, neutral, or negative tones. When applied to the agglomerated answers, the words included inside the positive grammar occurred 95 times (frequency: 95).

Conversely, the negative grammar (see Fig. 3) had a frequency of 45 tokens. Hence, the positive grammar weighs more than the negative. However, with words and expressions directly extracted from the answers, the striking element of this grammar is the students’ choice of vocabulary. The negative tokens are expressing rather intense negative sentiments (e.g., “disgusting”, “overkill”, “unimaginable”). When delving back into the students’ answers, most negative comments were made about (a) organisational aspects such as the food, the hotel, and the lack

of free time, and (b) one activity in particular, the rubbish collection. However, this last negative sentiment toward the activity reveals a potential positive insight. Indeed, such activity is often reluctant but has a solid environmental and anti-pollution message. Hence, we can imagine that despite the lack of enjoyment from the students, it indeed had a striking learning effect and potentially will implement more sustainable practices in their behavior.

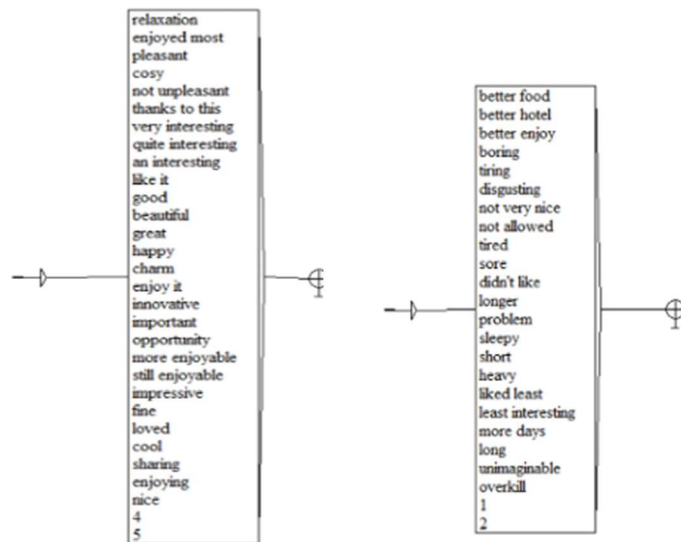


Figure 3. Positive and negative (from left to right) grammars for the Spanish students' answers.

#### 4.2.3. Co-occurrences analysis

Like the Portuguese students' answers, the Spanish answers confirm the following trends: The first half encompasses various peaks from negative and positive grammars, while the second half is nearly flat. These graph comparisons end with a negative peak. The empty parts refer to descriptive questions about the students' sustainability knowledge, in which they adopt a more neutral tone. The first part includes both peaks

due to questions about their trip satisfaction. They had nuanced answers, highlighting the trips' negative and positive aspects. The negative graph peaks at the end. In this case, the last question concerns their ideas for improvement, with which many of them express diverse criticism, as mentioned in the above paragraph. However, once again, positive words and expressions occurred 95 times versus 45 times for the negative ones. Quantitatively, the overall assessment of the pilot trip among Portuguese students is positive.

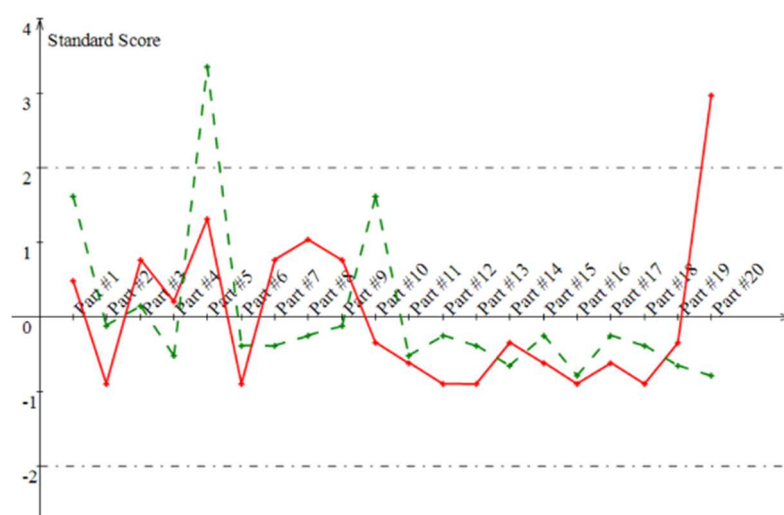


Figure 4. Graph comparison of the grammars positive (full) and negative (dash) in the Spanish students' answers.

### 4.3. Italian students

The fifteen survey respondents are 14-16-year-old students from Italy traveling to Spain.

#### 4.3.1. Key themes

The students often used words associated with the learning aspect of the sustainable pilot trip, such as “environment, sustainable, tourism”, and

they frequently discussed the organisational aspect of the trip and its activities with elements such as “trip, day, conferences, excursions”.

|             |          |             |         |             |            |
|-------------|----------|-------------|---------|-------------|------------|
| Tokens      | viaggio  | ambiente    | giornio | sostenibile | esempio    |
| Translation | trip     | environment | day     | sustainable | example    |
| Frequency   | 15       | 14          | 14      | 13          | 11         |
| Tokens      | attività | primo       | turismo | conferenze  | escursione |
| Translation | activity | first       | tourism | conference  | excursions |
| Frequency   | 9        | 9           | 9       | 7           | 7          |

Table 3. Most frequent relevant tokens.

#### *4.3.2. Sentiment analysis.*

Similarly to the previous responding groups (i.e., Spanish and Portuguese students), both positive and negative grammars were built on the tokens extracted from the answers. Some disambiguation was necessary to make sense of some words such as “tempi stretti” [meaning “short timing”, or literally “tight times”]; indeed, the words “times” and “tight” alone do not carry positive or negative values, whereas agglomerated together as written by a student, it carries a negative connotation. The negative grammar occurred 21 times (frequency: 21) and 113 times for the positive one (frequency: 113). The positive grammar is longer than any other grammar as for Italian, feminine, masculine and plural forms had to be added (e.g., “positiva” and “positivo”, “tipico” and “tipici”).

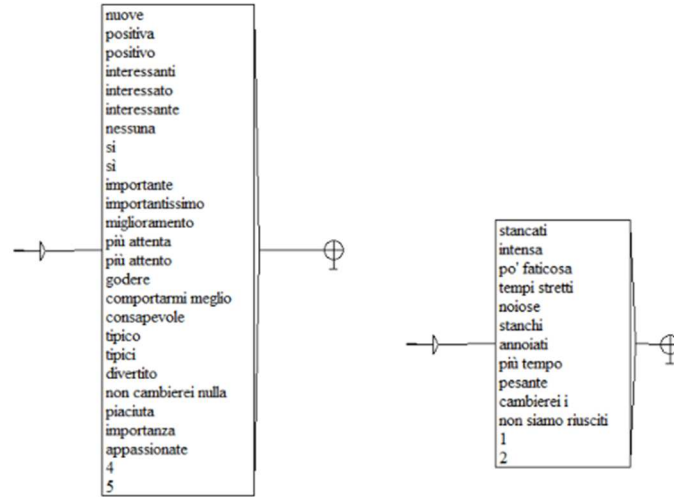


Figure 5. Positive and negative grammars (from left to right) for the Italian students' survey answers.

#### 4.3.3. Co-occurrences analysis

In the following graphs analysis, we can observe the highest peaks of both the negative and positive grammars occurring at the beginning. Indeed, the students are asked about their favorite and least favorite activities at the start of the survey. The most recurring positive comments refer to the outdoorsy character of the trip, such as “*andare al mare*” [“going to the sea”] and “*Il contatto con la natura e la biodiversità*” [“The contact with nature and biodiversity”] or to specific activities such as the excursion at Sierra de Tramuntana in Spain and the activity around the Posidonia. The most recurring negative comments refer to the conference held during the first day of the trip. Most students highlighted that this activity was tedious and that the first day was too intense, resulting in a general tiredness from the group. The positive grammar is peaking more than the negative one, which tends to be flat after the two peaks. Indeed, most survey answers refer to positive comments, especially concerning the learning outcomes. All students affirm learning about the importance of sustainable development and adopting responsible behaviors while traveling. The last red peak corresponds to

the suggestions for improvements. Some students recommended some improvement concerning the organisational aspects of the trip (choice of restaurants, free time allocation) and the selection of activities, such as fewer conferences, more activities at the sea and the beach, and about maritime sustainability.

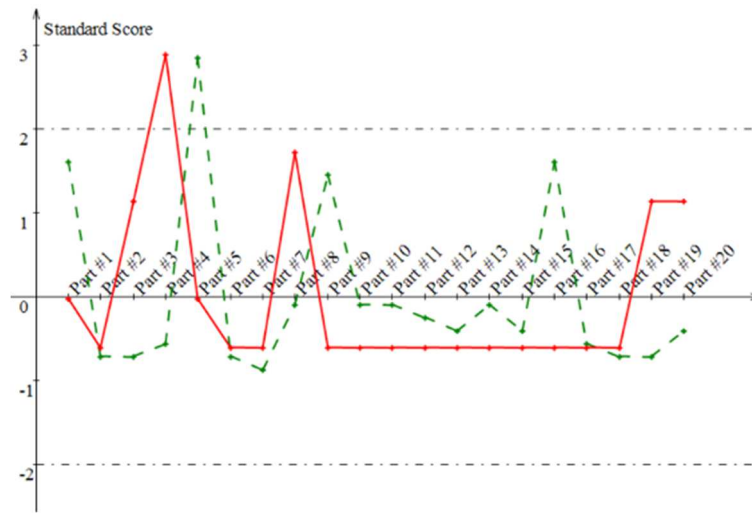


Fig. 6. Graph comparison of the positive (green) and negative (red) grammars in the Italian students' answers.

## 5. Conclusion

This article presented the analysis of the feedback from the students attending the WeMED pilot trips, which served as practical applications of the project's efforts, allowing participants to experience the eco-smart tourism packages first-hand.

The NooJ-supported analysis of the survey answers from the students in Portugal, Spain, and Italy reveals a complex interplay of posi-

tive and negative sentiments towards the pilot trips and their experiences. However, the generally positive sentiment surpasses the negative one. While most students demonstrated a gain of knowledge about sustainability, awareness of the environment, and a profound consciousness of sustainable tourism challenges, their overall satisfaction with the trip varied depending on several aspects. Key topics emerged from the data, including nature, environment, outdoor activities, and learning outcomes. For instance, one of the most mentioned words across the three student groups was “environment”. Some organisational aspects, such as activities, accommodations, and free time allocation, were subject to higher criticism. Positive sentiments revolved around engaging activities, knowledge growth, and a deepened understanding of sustainability challenges.

The findings highlight the rich potential of educational trips to engage students in sustainability awareness by putting it into practice (Činčera et al. 2022). Students learn through exposure activities (Bushell et al. 2011). Careful planning and organisation are crucial to the good conduct of these trips, which require the involvement of all stakeholders.

Further research should delve more deeply into the learning outcomes to grasp potential long-term learning for the students.

## References

- Ballantyne, Roy / Packer, Jan 2002. Nature-based Excursions: School Students' Perceptions of Learning in Natural Environments. *International Research in Geographical and Environmental Education* 11/3, 218–236.
- Behrendt, Marc / Franklin, Teresa 2014. A Review of Research on School Field Trips and Their Value in Education. *International Journal of Environmental and Science Education* 9, 235–245.

- Bihu, Reuben 2021. Questionnaire Survey Methodology in Educational and Social Science Studies. *International Journal of Quantitative and Qualitative Research Methods* 9/3, 40–60.
- Bushell, Brenda / Iwamura, Ryoko / Ozawa, Momoko / Goto, Masayuki 2011. Educating for sustainability: a pilot study in an elementary school in rural Nepal. *Interdisciplinary Environmental Review* 12/12, 12–23.
- Činčera, Jan / Kroufek, Roman / Bogner, Franz X. 2022. The perceived effect of environmental and sustainability education on environmental literacy of Czech teenagers. *Environmental Education Research* 29/9, 1276–1293.
- European Climate, Infrastructure and Environment Executive Agency, n.d. EU Commission Featured Project Page, <[https://cinea.ec.europa.eu/featured-projects/eu-wemednatour-ecotourism-packages-schools\\_en](https://cinea.ec.europa.eu/featured-projects/eu-wemednatour-ecotourism-packages-schools_en)> (accessed 05/09/2024)
- EU WeMED\_NaTOUR, n.d. About the Project Page <<https://www.euwemed-natour.eu/about-the-project>> (accessed 18/09/2024)
- Falk, John H. / Ballantyne, Roy / Packer, Jan / Benckendorff, Pierre 2012. Travel and Learning: A Neglected Tourism Research Area. *Annals of Tourism Research* 39/2, 908–927.
- George Washington University Libraries n.d. Text Analysis Methods Page, <<https://libguides.gwu.edu/textanalysis/methods#:~:text=Word%20frequency%20analysis%20in%20text,occurring%20words%20and%20their%20frequencies>> (accessed 20/08/2024)
- Monteleone, Mario / Mirto, Ignazio M. 2021. NooJ Grammars for Italian Negation System and Sentiment Analysis. In Bigey, Magali / Richeton, Annabel / Silberztein, Max / Thomas, Izabella (eds) *Formalizing Natural Languages: Applications to Natural Language Processing and Digital Humanities. NooJ 2021. Communications in Computer and Information Science*, vol. 1520, Cham: Springer, 39–50.
- Olivieri, Cecilia / Laquidara, Lorenzo M. / Semlali, Agathe 2024. Semantic Analysis of Migrants' Self-entrepreneurship Ecosystem Narratives. In Bartulović, Anita / Mijić, Linda / Silberztein, Max

(eds) *Formalizing Natural Languages: Applications to Natural Language Processing and Digital Humanities*. NooJ 2023. *Communications in Computer and Information Science*, vol. 1816. Cham: Springer, 176–187.

Silberstein, Max 2016. *Formalizing Natural Languages: The NooJ Approach*. London: Wiley-ISTE, London.

MARIO MONTELEONE

## Some Reflections on Italian Derivation and Computational Morphology

### Abstract

As known, Computational and Automatic Derivative Morphology routines are useful especially for speeding up the creation of exhaustive electronic dictionaries necessary to achieve complete and in-depth Natural Language Processing analyses. Furthermore, the structuring of these routines can be based on precise morphosyntactic and derivational rules, such as, for example, in Italian, the one that predicts, for each transitive verb, an adjectival derivation in *-bile*. However, it is worth noting that the Italian lexicon - together with many other lexicons - is not homogeneously structured: this means that its entries, as for their morphosyntax and semantic uses, are not all subject to the same rules of morphological derivation and segmentation. For this reason, a “wild card” application of Computational and Automatic Derivative Morphology routines, although well structured, is not possible.

Keywords: NLP, Natural Language Processing, Computational Morphology, Derivative Morphology, Morphological Segmentation, NooJ, NooJ Derivation Routines

## 1. Introduction

As known, Computational and Automatic Derivative Morphology (CADM) routines are useful especially for speeding up the creation of exhaustive electronic dictionaries<sup>1</sup> necessary to achieve complete and in-depth NLP analyses.<sup>2</sup> Furthermore, the structuring of these routines can be based on precise morphosyntactic and derivational rules, such as, for example, in Italian, the one that predicts, for each transitive verb, an adjectival derivation in *-bile*, as shown below:

*Paolo riscrive il documento =: Il documento è riscrivibile da Paolo*  
(Paolo rewrites the document =: \*?The document is rewritable by Paolo)

However, it is worth noting that the Italian lexicon - together with many other lexicons - is not homogeneously structured: this means that, as we will see, its entries, as for their morphosyntax and semantic uses, are not all subject to the same rules of morphological derivation and segmentation. For this reason, a “wild card” application of CADM routines, although well structured, is not possible. Unless derivation processes and routines are well contextualized. i.e. unless they come from solid and validated morphosyntactic and semantic observations, if applied automatically they can produce imprecise results, mainly in terms of signifiers with non-existent significances<sup>3</sup>, and for some classes of verbs, in terms of non-existent or inapplicable morphosyntactic and semantic properties. Ultimately, these faulty procedures and their resulting errors can invalidate the efficacy of a large number of NLP analyses.

---

<sup>1</sup> See Gross, Maurice, 1989a, 1989b and 1995.

<sup>2</sup> One of the main methods of application of CADM routines that we refer to here concerns the DRV (derivation) command available in the NooJ NLP environment for the creation and applications of electronic dictionaries. For further information, it is possible to consult the titles indicated in the references, relating to this environment, as well as the further examples provided in this paper.

<sup>3</sup> As for this mismatch, the philological features of some entries also plays an important role.

For instance, as for the aforementioned derivation of adjectives in *-bile*, we will see how some of them can originate from intransitive verbs, as *passaggiabile* (walkable) from *passaggiare* (to stroll, intransitive in Italian) in sentences as:

*Il sentiero è facilmente passeggiabile* (The path is easily walkable)

Consequently, in our study, we will try to establish if, when and how CADM routines can be structured and applied, in Italian, via wildcard commands. We will try to establish when and how apply automatic morphological derivations based on word lexical uses and – for predicative elements – syntactic profiles. The word classes we will include in such analysis are:

- Adjectives in *-bile*;
- Iterative and pronominal verbs;
- Mode adverbs in *-mente*, as *milanesemente* (in the Milanese way) derived from *Milanese*, that is from all adjectives derived from Italian city proper nouns;
- As for Part of Speech Migration, Proper Names of City Administrative Areas, Exclamations and Onomatopoeia.

## 2. Some Brief Considerations on Derivational Morphology

In natural languages, the application of Derivational Morphology (DM) mechanisms can result in both words that are morphologically different from the initial elements, and in words that acquire, in addition to a morphological diversity, also specific syntactic differences.

For instance, as already stated, in Italian the mode adverbs ending in *-mente* are in large part derived by adding this suffix to a feminine singular adjective. In such cases, we will have two words sharing the same lexical morpheme but having different uses. Instead,

iterative and pronominal verbs derive, through specific suffixation, from non-iterative and non-pronominal verbs, compared to which they can have different syntactic behaviours, hence can occur in sentences characterized by different syntactic contexts. In such cases, we will have verbs having contiguous meanings but different distributions and (acceptable and grammatical) transformations.

However, the morphosyntactic features of many natural languages – Italian included – do not seem predisposed to allow wildcard applications of all their DM derivative procedures to all their lexical morphemes. Therefore, the question we should here try to answer is: what appropriateness would there be in developing a Natural Language Processing (NLP) system, or rather, a Computational Morphology (CM) system that automatically derives words from each other, without taking into account all types of applicable restrictions, from semantic to logical-deductive ones?

As we will see in the following pages, the margins of appropriateness of such a procedure would be – and are – concretely questionable. In fact, very high is the risk of producing, even for native speakers of a given language, words with improbable or non-existent meaning, difficult to pronounce, with impossible or bizarre uses, and with non-acceptable and non-grammatical morphosyntax. In short, the risk would be that of not respecting, in any aspect, all the Saussurean dichotomies, and in particular that relating to the dynamics existing between *Langue* and *Parole*.

It is worth stressing that with such assertions we do not want to cast doubt on all NLP procedures and CM ones. On the contrary, we simply want to depict and exemplify a morphosyntactic derivation system that is effective, efficient, and linguistically sustainable. These purposes are therefore important and crucial, since for many linguistic families and natural language, DM is one of the most discussed topics, from both a General Linguistics (GL) and an NLP point of view.

### *2.1. Origins of Modern Derivation Morphology and Morphological Segmentation*

As known, modern DM bases its mechanisms and functioning on the method of word segmentation into morphemes first experimented and perfected on Amerindian languages by the American linguists starting from 19th century<sup>4</sup> and of the first years of the 20th (i.e. Franz Boas, Edward Sapir, and Benjamin Whorf). Regarding these languages, non-native speakers had to recur to segmentation in order to understand until then unknown morphosyntactic and semantic structures. In fact, Amerindian languages are polysynthetic, and use “sentence words”,<sup>5</sup> in which thanks to specific grammar cases each morpheme can have both morphological and syntactic value.

We also stress that at the historical moment of their interest in the Amerindian languages, 19<sup>th</sup> century linguists had two main concerns. The first was to assuage the common sense of guilt arising within a large part of the American public opinion for having violently brought Amerindians very close to extinction before and during the Indian Wars. The second was not to lose Amerindian linguistic heritage<sup>6</sup>, which risked disappearing, in its enormous variety, together with its speakers. Therefore, to allay this last concern, it was necessary both to dialogue with the last remaining Amerindians and to try to understand how their polysynthetic languages were structured and used.

Later on, segmentation into morphemes structured and practiced by American linguists was applied to other linguistic fields and other analytical levels. Today segmentation into immediate constituents is commonly used to account for the morphology and syntax of many languages, even by means of different and sometimes antithetical analytical approaches. This is shown, for instance, by the birth and development of both Structuralist Formal and Computational

---

<sup>4</sup> See Boas 2013 and Campbell 1997.

<sup>5</sup> See for instance [https://en.wiktionary.org/wiki/Category:Lakota\\_language](https://en.wiktionary.org/wiki/Category:Lakota_language).

<sup>6</sup> We underline, by way of example, that a large part of the Toponymy of the Americas is of Amerindian origin (see Monteleone 1986) such as for example the word Niagara, which derives from the Iroquois Onguiaahra, translatable as “waters that fall rumbling”.

Syntax, and Generative Linguistics. In such analyses, since its first applications, segmentation, and subsequently DM gave the impression of being suitable for wildcard applications, thanks to the (presumed and unprovable) systematicity of their results. This assumption affected somehow CM automatisms, especially when producing high numbers of false positives. In any case, a more detailed analysis of the various lexicons has shown that especially for those languages having a complex inflection system, segmentation and DM wildcard applications do not always produce valid results. In fact, as already stated, if we closely examine for instance, Italian lexicon, as we will do in the following pages, we conclude that it is not structured: by means of segmentation, Italian words, in their various forms and formations, do not allow for deducing and/or establish general and uniformly valid morphological formation and derivation rules.

These substantiations essentially demonstrate two distinct aspects:

1. As established by Lexicon-Grammar<sup>7</sup> for all predication elements (verbal, nominal and adjectival), also the rules of word formation must be contextualized from a morphosyntactic and semantic point of view;
2. The method of defining word formation features can only be a bottom-up one, i.e. it has to be based on the factual study of linguistic data.

Therefore, it ensues that segmentation into morphemes, and subsequently DM, together with CM procedures, must be carried out based on all the characteristics of the words examined or to form, namely: morphological, syntactic and semantic. With such premises, automatisms do not seem to be the best solution applicable.

---

<sup>7</sup>

<https://en.wikipedia.org/wiki/Lexicon-grammar>.

## *2.2. NooJ Derivation Routine*

The NooJ<sup>8</sup> Derivation routine (DRV) can be taken as an example of automatic morphological derivation tool that needs to be used *cum grano salis*. As a rule-based NLP environment, NooJ provides for the creation, morphosyntactic annotation and inflection of electronic dictionaries<sup>9</sup>, which are applicable to manage various aspects of automatic textual analysis, as well as knowledge automatic representation through localizations, extractions and formalizations of concepts. Besides, Silberztein (2016, page 100) states that:

An electronic dictionary must link each lexical entry to its derived forms. Forms derived from an ALU (Atomic Linguistic Units) must not be described separately from one another. Therefore, it would be absolutely necessary to add to the electronic dictionary a formalization of derivations [...] For example, we must have a single lexical entry to describe the verb to demonstrate, the abstract noun a demonstration and the human noun a demonstrator. In this way, we can link the sentences: *There were a lot of people demonstrating against the war* = *There were a lot of demonstrators against the war* = *A lot of people had a demonstration against the war*.

However, concerning Italian language, as already mentioned and as we will see in the following pages, a "wild card" application of formal word derivation rules is risky, both because it can produce several false positives and also because it will not account for words that are connected to each other only formally, but not semantically. Even though we do not have here the purpose to question the validity of DRV NooJ routine, we want to stress that its applications must be carefully calibrated, with reference to any lexicon, avoiding a priori solutions. In this sense, the DRV NooJ routine should be used as a tool to create lexical sets with contiguous morphosyntactic features and meanings, avoiding extensive applications that may produce, as mentioned, false positives (and we write it again, well aware of appearing repetitive).

---

<sup>8</sup> <https://en.wikipedia.org/wiki/NooJ>.

<sup>9</sup> See Gross 1989a, 1989b and 1995.

### 3. Some Examples of Inaccurate Automatic Morphological Derivations for Italian Adjectives in *-bile*

In Italian<sup>10</sup>, the suffix *-bile* is added to lexical morphemes to form adjectives. The meaning that this suffix gives to the adjective it forms is generally rendered with the periphrasis “which can be x”, where x is the past participle of the verb from which the adjective derives. Commonly, it is said that the formation of these adjectives is possible only applying a passive transformation to declarative sentences in which the operator is a transitive verb. For instance:

*Max beve l'acqua* = [Passive] *L'acqua è bevuta da Max* = [ADJ -> *bile*] *L'acqua* + *bevibile* [E + *da Max*]  
(Max drinks water = Water is drunk by Max = Water is drinkable [E+by Max])

This derivation procedure is very productive, therefore the number of Italian adjectives in *-bile* is very high (theoretically, one for each transitive verb). As already stated, since “to be + past participle” is a passive verb form, this type of adjective is supposed to be not derivable from all verbs, but only from the transitive ones, which admit the passive. For the same reason, any adjective in *-bile* should always be completed by what logical analysis indicates as the complement of agent and/or efficient cause, introduced by the preposition *da* (by).

However, we state here and demonstrate using the previous derivation model, it is possible to create also those *-bile* adjectives not attested in paper dictionaries, or in principle not allowed by the transitive-derivation model itself.

Moreover, the regularity of the transitive-verb morphologically derivation procedure would seem to effectively conform to the automatisms offered by the NooJ DRV function [1]. However, in creating the NooJ Italian Dictionary of Simple Words (DELAS), we

---

<sup>10</sup> See <https://accademiadellacrusca.it/it/consulenza/immettibile-o-immissibile-non-si-pu-mettere-o-si-pu-immettere/13123>.

observed two particularities, which, in fact, make the aforementioned regularity of derivation imprecise:

- The are some Italian adjectives in *-bile* that are not derived by a verb, as:
- *potabile* (drinkable), which cannot be derived from *potare* (to prune, to trim);
  - *papabile* (from *papa* (pope), meaning suitable, sellable, who can be elected), considering that in Italian we have no verb *\*papare*.

In addition, some Italian intransitive verbs (i.e. those accepting passive or causative transformations) may derivate adjectives in *-bile*, as:

- *fruibile* (from *fruire*, intr. (to use, to make use of), widely used in in bureaucratic language);
- *marciabile* (from *marciare*, intr. (to march), also with reference to a piece of music that can be played in marching rhythm); the opposite *immarciabile* is also attested;
- *passaggiabile* (from *passaggiare*, intr. (to walk, to stroll), with reference to a route easily passable on foot); the opposite *impassaggiabile* is also attested;
- *rinunciabile* (from *rinunciare*, intr. (to renounce), renounceable); the opposite *irrinunciabile* is also attested.

We also stress in Italian the use of the fixed productive structure "*V tutto il ADJ-bile*", as in, which is also valid for some intransitive verbs such as *marciare* (to march), or *ponderare* (to ponder), which has a widely employed intransitive use, that is, *ponderare su* (to ponder about/on):

- *mangiare tutto il mangiabile* (to eat all can be eaten);
- *ponderare tutto il ponderabile* (to ponder all that is ponderable);
- *passaggiare tutto il passaggio* (walk around the entire walkable area);
- *marciare tutto il marciabile* (march all the marchable).

Among Italian *-bile* adjectives produce by intransitive verbs, only as concerns the letter “a” of Italian lexicon, we can find: *accessibile* (accessible), *acconsentibile* and *adempibile* (fulfilable), *affacciabile* (faceable), *affiorabile* (surfaceable), *affogabile*, (drownable), *agibile* (usable), *allegoreggiabile*, (allegorable), *alloggiabile* (accommodateable), *allineabile*, (alignable).

This demonstrates that the rule of morphological derivation based exclusively on the transitivity of verbs is only partially precise.

#### 4. Some Examples of Inaccurate Automatic Morphological Derivations with Italian Prefixed and Suffixed Verbs

Here we will examine the verb forms obtainable through iterative prefixations (IPs) and pronominal suffixations (PSs). For instance:

- *rilanciare* (*ri+lanciare*; to relaunch, to restart; to reintroduce, to revamp; to raise);
- *lanciarsi* (*lanciar+si*; to jump, to dive, to launch/to haul/to throw oneself).

For many of these verbs, an automatic processing of morphological derivation can often produce false positives, i.e. analytical IP or PS verbs that, with reference to the word or words from which they derive, have autonomous morphosyntactic and semantic profiles.

Here, we will now show how to structure automatic IP and PS to avoid producing false positives. In order to do so, it is necessary to carry out a careful morphosyntactic contextualization of each verb subjected to IP and PS, verifying the results of the two derivation procedures.

#### *4.1. Morphosyntax of Iterative and Pronominal Prefixation*

We note that in most cases, in a Lexicon-Grammar perspective,<sup>11</sup> IP may be seen as an unnecessary synthetic routine (manly a stylistic one), easily replaceable with the corresponding analytical structures. In fact, Italian iterative verbs obtained with the prefix *-ri* placed at the beginning of given infinitives should not be used as ALUs of electronic dictionary, as their occurrence do not produce simple sentences. In fact, they can be transformed into sentences containing metadiscursive (hence, erasable) adverbs. For instance:

*Paolo riscrive la sceneggiatura* =: *Paolo scrive la sceneggiatura di nuovo*  
(Paolo rewrites the script = Paolo writes the script again)  
*Paolo riracconta la barzelletta* =: *Paolo racconta la barzelletta nuovamente*  
(Paolo retells the joke =: Paolo tells the joke again)

Besides, some Italian iterative verbs can also have different, additional, and autonomous meanings, as for instance:

- *rilevare* (to detect), which does not only mean *levare di nuovo* (to take off again);
- *rifutare* (to refuse), which does not only mean *fiutare di nuovo* (to smell again);
- *rispettare* (to respect), which does not only mean *spettare di nuovo* (to be up again to).

As for pronominal verbs obtainable through PS, we note that in Italian we have the following pronominal verb typology, deriving mainly from well-rooted lexicalization and syntax procedures. Italian verbs ending in *-rsi* can be:

- reflexives, as *pettinarsi* (to comb oneself) in *Max si pettina* (Max combs himself);
- possessive reflexives, for instance:
  - *lavarsi Nparte\_del\_corpo* (to wash Nbody\_part), as in *Max si lava le mani* (Max washes his hands);

---

<sup>11</sup> See Boons, Guillet, Leclère 1976a and 1976b.

- *abbottonarsi Nabito* (to button one's Ndress), as in *Max si abbottona la camicia* (Max buttons his shirt);
- pronominal intransitives as:
  - *ammantarsi* (to be mantled with) in *La collina si ammanta di neve* (The hill is mantled with snow), in a binary transformation relationship with *La neve ammanta la collina* (The snow mantles the hill);
  - *inginocchiarsi* (to kneel), which cannot be derived from the transitive verb *\*inginocchiare*;<sup>12</sup>
- pronominal transitives as *accaparrarsi* (to grab) in *Harry si accaparra l'uovo doro* (Harry grabs the golden egg), which may be paraphrased into *Harry accaparra l'uovo d'oro per se stesso*;
- reciprocals such as *sposarsi* (to get married) in sentences as *Max e Paola si sposano* (Max and Paola get married).

It is important to underline that, a part from the two types of pronominal transitives, all the verb falling into the previous typology are not to be considered ALUs, as they can be syntactically obtained through prefixation of canonical infinitive forms (*pettinare, lavare, abbottonare, ammantare, accaparrare, sposare*).

It ensues that, apart from pronominal transitives, the addition of the NopoJ tag "DRV=RSI" for any other Italian verb can lead to the generation of false positives. In such cases, we generate verbs that are not autonomous pronominal forms such as for example *inginocchiarsi*, but are syntactically derived from non-pronominal verbs, with which they share only the same lexical morpheme.

As for these restrictions, some few examples are:

- *portare* -> *portarsi* = *portare se stesso* = Reflexive, as in: *Il giocatore si porta sulla fascia destra del campo* (The player moves to the right side of the field);
- *lanciare* -> *lanciarsi* = *lanciare se stesso* = Reflexive, as in *Paolo si lanciò sulla frittata* (Paolo threw himself on the omelette);

---

<sup>12</sup> In Italian, this inflected form of *inginocchiarsi* is used only in causative constructions as *Paolo fa inginocchiare Marco* (Paolo makes Marco kneel).

- *bruciare* -> *bruciarsi* = *bruciare* DET ADJ\_Pos\_N0 = Possessive reflexive, as in *Paolo si bruciò l'occasione* (Paolo missed the opportunity);
- *impegnare* -> *impegnarsi* = *impegnare* DET ADJ\_Pos\_N0 = Possessive reflexive, as in *Paolo si è impegnato la casa* (Paolo has pawned his house).

Besides, as already mentioned, we note that many Italian iterative verbs cannot be considered electronic dictionary entries, as when used as operators, they do not produce simple sentences but sentences containing metadiscursive (hence, erasable) adverbs. For instance:

*Max riscrive la sceneggiatura* = *Max scrive la sceneggiatura di nuovo* (Max rewrites the script = Max writes the script again)  
*Max riracconta la barzelletta* = *Max racconta la barzelletta nuovamente*  
(Max retells the joke = Max tells the joke again)

In such case, a wildcard application to verbs of the NooJ notation “DRV=RI” (or RE, depending on the starting vowel of the verbs) would produce not only (few) usable ALUs, but also false positives, of which we give here only some of the many examples:

- *riporre* (to save, to store, to store away) ≠ *porre di nuovo* (to put again, to place again, to layagain; to set again, to establish again, to fix again; to suppose again, to imagine again);
- *rilevare* (to detect) ≠ *levare di nuovo* (to take off again);
- *rifutare* (to refuse) ≠ *fiutare di nuovo* (to smell again);
- *rispettare* (to respect) ≠ *spettare di nuovo* (to be up again to);
- *risorgere* (to resurrect) ≠ *sorgere di nuovo* (to rise; to arise, to appear, to show up; to rise up, to stand).

## 5. Some Examples of Inaccurate Automatic Morphological Derivations with Italian Mode Adverbs

As known, in Italian, mode adverbs in *-mente* are formed starting from feminine adjective singular forms, to which the suffix is added, For instance:

- *automatica-mente* (automatically)
- *brusca-mente* (abruptly)
- *semantica-mente* (semantically)

To form their corresponding mode adverb, the adjectives ending in *-le*, drop the final *-e*, as in:

- *comprensibile* -> *comprensibil-mente* (comprehensibly)
- *elettorale* -> *elettoral-mente* (electorally)
- *mentale* -> *mental-mente* (mentally)

It should not be forgotten that all Italian adverbs in *-mente* are synthetic forms that can be paraphrased with *in modo* (in a way) ADJ, *in maniera* (in a way) ADJ, or *da un punto di vista* (from a point of view) ADJ. In these derivations, the ADJ is the same adjective used for the generation of the adverb, and grammatically agreed to the noun used in the expression:

- *automatica-mente* (automatically) = *in modo automatico/in maniera automatica*
- *brusca-mente* (abruptly) = *in modo brusco/in maniera brusca*
- *semantica-mente* (semantically) = *da un punto di vista semantico*

We can therefore state that the generation and use of many adverbs in *-mente* can also be considered as a stylistic procedure, morphosyntactically not necessary. Besides, in some cases and with specific classes of adjectives, the adverbs produced suffer from cacophony, or bear a meaning that is not entirely understandable, due to their very rare use. In such cases, the block to the generation of a

given adverb may seem essentially, due to semantic or pragmatic acceptability.

This implies that although the formation of *-mente* adverbs has regular derivative procedures, for Italian electronic dictionaries it does not seem possible to create an automatic routine from them through, for example, a NooJ command of the type “DRV=MENTE”. For example, in the class of adjectives that refer to city names and inhabitants, in the following list we find cases that do not seem easily usable, notwithstanding the regularity of the DM procedure:

- *napoletanamente* (the way they do it in Naples) = *alla (E+maniera) napoletana*
- *romanamente* (the way they do it in Rome) = *alla (E+maniera) romana*
- *fiorentinamente* (the way they do it in Florence) = *alla (E+maniera) fiorentina*
- *pisanamente* (the way they do it in Pisa) = *alla (E+maniera) pisana*
- *salernitanamente* (the way they do it in Salerno) = *alla (E+maniera) salernitana*
- *\*?bergamascamente* (the way they do it in Bergamo) = *alla (E+maniera) bergamasca*
- *\*?bolognesemente* (the way they do it in Bologna) = *alla (E+maniera) bolognese*
- *\*?milanesemente* (the way they do it in Milan) = *alla (E+maniera) milanese*
- *\*?genovesemente* (the way they do it in Genova) = *alla (E+maniera) genovese*
- *\*?torinesemente* (the way they do it in Turin) = *alla (E+maniera) torinese*
- *\*?leccesemente* (the way they do it in Lecce) = *alla (E+maniera) leccese*

## 6. Some Examples of Inaccurate Automatic Morphological Derivations with Italian Part of Speech Migration

### 6.1. Proper Names of Administrative Areas

In Italian, all city proper names generates a specific adjective, which in its masculine singular form can be used as a proper noun indicating the specific geographical zone and area a given city administers.

Although this procedure may appear to be regular, therefore easy to formalize through the DM and a NooJ DRV routine, it is necessary to underline how, again for acceptability but also philological reasons, in several cases, an automatic part of speech (POS) migration could present some issues.

For instance, in the following list, we note how the use of some of these forms is blocked or varied, in spite of the regularity of the formation procedure:

- *nel Napoletano* = in the areas administered by the City of Naples
- *\*nel Romano* = in the areas administered by the City of Rome
- *nel Salernitano* = in the areas administered by the City of Salerno
- *nel Bergamasco* + *nella Bergamasca* = in the areas administered by the City of Bergamo

Furthermore, this POS migration is less common (i.e. less attested) for the proper names of cities that are not famous, or that are not provincial capitals:

- *\*nell'Abbateese* + *nell'Abbateese* = in the areas administered by the City of Abbiategrasso;
- *\*?nel Caltanissetese* + *nel Nisseno* = in the areas administered by the City of Caltanissetta;
- *?nel Trevisano* + *nel (Trevigiano+Trivigiano)* = in the areas administered by the City of Treviso.

## 6.2. Exclamations

Based on what is indicated by classical lexicology and lexicography, exclamation is a POS that may incorporate uses and words migrated from various other PsOS. Anyway, in the case of verbs migration towards exclamation POS, it is worth stressing that this always produce metadiscursive verb sentences. For example:

*(Io dico che lui esclamò:) “Vattene!!!”* ((I say he exclaimed:) “Go away!!!”)  
*(Io dico che lui esclamò:) “Basta!!!”* ((I say he exclaimed:) “That’s enough!!!”)

In such cases, it will therefore be imprecise to refer to these expressions as exclamations, and it will instead be correct to define them as reductions of simple or elementary sentences.

Also onomatopoeias are regularly included among exclamations. However, it must be underlined that onomatopoeias should be ascribed as autonomous POS, and that in many cases they can be predicative elements in Verb Support constructions.<sup>13</sup>

*Il gatto miagolò = Il gatto fece miao* (The cat meowed = The cat made a meow)  
*Paolo grugnì = Paolo fece grunt* (Paolo groned = Paolo let out a grunt)  
*La porta scricchiolò = La porta fece crack* (The door creaked = The door made a crack)

## 7. Conclusions and Further Research

At the end of this study, we want to remark that for the sake of brevity, we have not exposed all the cases in which, in Italian, a wild card DM application might produce mainly false positive, hence being prone to a deep and accurate revision. This is the case, for example, with all forms of alteration, such as the adjectives in *-issimo* (absolute superlative) producing nouns which cannot be reduced to standard

---

<sup>13</sup> See Boons, Guillet, Leclère 1976a, 1976b, and Gross 1975.

form. This is the case, for instance, of the feminine word *direttissima* (immediate) in the compound word *processo per direttissima* (immediate trial), which does not mean *molto diretta* (very direct).

Therefore, concerning the topics exposed in this study, in a near future we intend to expand the sets of words to which the previously exposed formal and morphosyntactic analyses can be applied, in order to categorise effective derivative rules to be included as ALUs, for instance, in Italian NooJ electronic dictionaries.

Anyway, as mentioned before, this study does not want to undermine either the value or the functionality of NooJ DRV tag and procedure. Rather, the aim is to highlight, always from a Lexicon-Grammar perspective that it is of crucial importance to contextualize every type of "manipulative" operation on natural language, pragmatically, and verifying its premises and results. This means that also NooJ DRV routines can be applied via wildcard commands, but only after having well established all derivation rules, even those applicable to just one word or ALU.

Besides, it would seem useful to contemplate the possibility to extend the use of NooJ DRV tag also to other forms of derivation, namely syntactic-semantic ones, and more specifically in all cases when:

- A syntactic transformation is blocked by a specific syntactic properties of a given verb operator;
- Nonetheless, it is possible to produce systematically the meaning of the above blocked transformation using another verb, with a contiguous meaning, thus obtaining a semantic/ontological classification of verb. For example:

*Il tempo fa diventare grande Max ≠ Il tempo fa ingrandire Max*

*Il tempo fa diventare grande Max = Il tempo fa crescere Max*

(Time makes Max grow up ≠ Time makes Max bigger = Time makes Max grow)

## References

- Boas, Franz 2013 (ed). *Handbook of American Indian Languages*. Cambridge University Press:Cambridge.
- Boons, Jean-Paul; Guillet, Alain; Leclère, Christian 1976a. *La structure des phrases simples en français. Constructions intransitives*. Librairie Droz :Genève-Paris, 377 p.
- Boons, Jean-Paul; Guillet, Alain; Leclère, Christian 1976b. *La structure des phrases simples en français. II:Classes de constructions transitives*. Technical report, Univ. Paris 7:Paris.
- Campbell, Lyle 1997. *American Indian Languages*. Oxford University Press:Oxford.
- Gross, Maurice 1975. *Méthodes en syntaxe*. Hermann:Paris.
- Gross, Maurice 1989a. La construction de dictionnaires électroniques du français. Traitement des langues naturelles. Ecole d'été du CENT. p. 5-36.
- Gross, Maurice 1989b. La construction de dictionnaires électroniques. *Annales des Télécommunications*, tome 44, n° 1-2, p. 4-19, Issy-les-Moulineaux/ Lannion:CNET.
- Gross, Maurice 1995. Quelques considérations sur les marques sémantiques d'un dictionnaire électronique. *Lexiques-grammaires comparés et traitements automatiques*, Actes du deuxième colloque international LGC, UQAM:Montréal, p.9-30.
- Grossmann, Maria; Rainer, Franz (edd.) 2004. *La formazione delle parole in italiano*. Tübingen:Max Niemeyer Verlag.
- <https://accademiadellacrusca.it/it/consulenza/fruibile-da-o-fruibile-a-e-se-fosse-fruibile-per/990>.
- [https://en.wikipedia.org/wiki/American\\_Indian\\_Wars](https://en.wikipedia.org/wiki/American_Indian_Wars).
- [https://en.wikipedia.org/wiki/Genocide\\_of\\_indigenous\\_peoples](https://en.wikipedia.org/wiki/Genocide_of_indigenous_peoples).
- [https://en.wiktionary.org/wiki/Category:Lakota\\_language](https://en.wiktionary.org/wiki/Category:Lakota_language).
- Lakota Language Consortium, <https://lakhota.org/introduction/>
- Monteleone, Mario 1986. Amerindianismi. Prestiti e calchi linguistici nell' inglese e nel francese del Canada. In AA. VV., *Annali della Facoltà di Lingue e Letterature straniere, 3a serie, VII, 1-2*, Bari, p. 27-48.

- Monteleone, Mario 2024. NooJ Grammars for Morphophonemic Continuity and Semantic Discontinuity. *Formalizing Natural Languages: Applications to Natural Language Processing and Digital Humanities*, Springer Nature Switzerland:Cham.
- Omniglot, <https://www.omniglot.com/language/phrases/lakota.php>.
- Silberstein, Max 2016. *Formalizing Natural Languages: The NooJ Approach*. London:ISTE Ltd and Wiley.