Macrothink Institute™

International Journal of Linguistics
ISSN 1948-5425
2025, Vol. 17, No. 6

From ‘il’ to ‘uno’: A Developmental Scale for Acquisition of Italian Article in L2

Anna Mantovani

University of Padova, DiSLL

Via E. Vendramini, 13, 35137, Padova (PD), Italy

E-mail: anna.mantovani.1@phd.unipd.it

Received: July 24, 2025 Accepted: August 28, 2025 Published: October 9, 2025

doi:10.5296/ijl.v17i6.23209 URL: https://doi.org/10.5296/ijl.v17i6.23209

Abstract

This study investigates the developmental trajectory of article acquisition in adolescent learners of Italian as a second language (L2), based on a corpus of written production from 208 students representing 35 different L1 backgrounds. Drawing on markedness theory (Jakobson, 1968; Eckman, 1985) and frequency-based approaches (Ellis, 2002), the analysis examines whether learners initially acquire unmarked and high-frequency forms and how morphophonological complexity influences accuracy and emergence. This study employs a quantitative, time series analysis (Xu, 2023) to investigate whether a developmental trajectory of article emergence and stabilization can be empirically observed over time. Learners’ productions were timestamped and processed, enabling the construction of longitudinal frequency matrices and the application of rolling averages. The stabilization of each article form was operationalized, allowing the reconstruction of an acquisition sequence. Findings support a frequency- and markedness-driven developmental hierarchy of Italian article acquisition shaped by both frequency-driven and phonological constraints. More marked forms tend to stabilize later and are often avoided in favor of less marked alternatives. Definite articles *il*, *la*, *i*, *le*, and indefinite *un* are among the earliest to stabilize and are frequently overused across all L1 groups, with a notable preference for feminine forms. Statistical analyses, including ANOVA and post hoc tests, confirmed significant variation in article use over time and between L1 groups.

Keywords: Italian as a second language, Article acquisition, Acquisition sequences, Developmental stages, Markedness theory, Time series analysis

1. Introduction

1.1 Problem Statement

The acquisition of articles in Italian as a second language (L2) remains a challenging area, as it involves phonological, morphosyntactic, semantic, and discourse-pragmatic aspects. Although several models have been proposed to explain the grammaticalization of articles, few studies have systematically explored how these frameworks apply to the learning of Italian articles by L2 learners from typologically diverse first language (L1) backgrounds. Furthermore, a critical yet underexplored aspect of this process is the role of markedness. Trubetzkoy (1939), cited in Bybee (2010, p. 132), defined markedness as a privative (binary) opposition, where the member with a 'mark' is considered the marked member, while the unmarked member is characterized by its absence. Later, Greenberg (1966) suggested that frequency is related to markedness: the unmarked member of an opposition tends to occur more often or with higher frequency in actual language use. This implies that article forms that are both unmarked and frequent in the input may be acquired earlier and more consistently than their marked, less frequent counterparts. The perspective aligns with the Markedness Differential Hypothesis (MDH) (Eckman, 1985, p. 296), which posits two key principles: (a) less marked structures are acquired earlier than more marked ones; (b) generalization tends to flow from marked to unmarked forms. Accordingly, Italian article acquisition may be constrained not only by input frequency or cognitive processing demands, but also by typological hierarchies that classify certain forms as more or less marked. Despite these theoretical insights, a comprehensive analysis integrating frequency, markedness, and temporal stability across multiple L1s has so far remained scarce in the context of Italian L2 acquisition. Therefore, in line with Giacalone Ramat's (2008, p. 13) observation that frequency alone does not account for developmental patterns, the present study investigates whether a developmental sequence informed by both frequency and markedness can be identified in L2 learners' production of Italian articles.

1.2 Grammaticalization Pathway of Italian Articles

Italian articles vary according to gender, number, and phonological context, and are phonologically cliticized to the following word (Giusti, 1997). The Italian articles' system is morphophonologically rich: the selection of the appropriate article thus requires the specification of grammatical features, namely the gender and number of the noun, and phonological features, particularly the onset of the following lexical item.

(1) un libro
a.INDF.M.SG book.M.SG
'a book'

(2) una casa
a.INDF.F.SG house.F.SG
'a house'

(3) un' amica
a.INDF.F.SG friend.F.SG

‘a girlfriend’

(4) uno scoiattolo
a.INDF.M.SG squirrel.M.SG
‘a squirrel’

(5) il libro
the.DEF.M.SG book.M.SG
‘the book’

(6) la casa
the.DEF.F.SG house.F.SG
‘the house’

(7) l’ amico
the.DEF.M.SG friend.M.SG
‘the friend’

(8) l’ amica
the.DEF.F.SG girlfriend.F.SG
‘the girlfriend’

(9) lo scoiattolo
the.DEF.M.SG squirrel.M.SG
‘the squirrel’

(10) i libri
the.DEF.M.PL books.M.PL
‘the books’

(11) gli scoiattoli
the.DEF.M.PL squirrels.M.PL
‘the squirrels’

(12) le case
the.DEF.F.PL houses.F.PL
‘the houses’

(13) mia madre
my.1SG.POSS mother.F.SG
‘my mother’

(14) Anna
Anna.F.SG
‘Anna’

The indefinite article in Italian shares its form with the numeral *uno* (‘one’) and is generally used to introduce a new referent in the discourse (Dal Pozzo, 2022, p. 357). The default masculine singular form is INDF.M.SG.*un* (1), which appears before consonants other than /s/+consonant clusters or /z/. The feminine singular form is INDF.F.SG.*una* (2), while INDF.F.SG.*un’* (3) is used before vowel-initial feminine singular nouns. A marked masculine singular form, INDF.M.SG.*uno* (4), appears before masculine nouns beginning with /s/+consonant clusters, /z/, and other sonority-violating onsets. In contrast, definite articles encode referents that are assumed to be familiar or identifiable within the discourse model (Ward & Birner, 1995, p. 726). The default masculine singular form is DEF.M.SG.*il* used before consonants with a simple onset (5) or clusters consisting of an obstruent + liquid/glide. The feminine singular form is DEF.F.SG.*la* (6), while the elided form DEF.SG.*l’* is used for both vowel-initial masculine (7) and feminine (8) singular nouns. The marked masculine singular form DEF.M.SG.*lo* is used before masculine nouns with initial /s/+consonant clusters, /z/, /gn/, /ps/, and similar onsets (Repetti, 2020), as in (9). In the plural, Italian uses DEF.M.PL.*i* for masculine nouns with simple onsets (10), and DEF.M.PL.*gli* as the plural counterpart of DEF.M.SG.*lo* and DEF.M.SG.*l’* (11). Feminine plural nouns take DEF.F.PL.*le* (12).

In standard modern Italian, the definite article is omitted in specific syntactic environments, such as with singular nouns denoting close family relations (13) unless modified (for instance, la mia cara madre, ‘[the] my dear mother’), and before personal names (14) unless accompanied by an attribute (e.g., la giovane Anna, ‘the young Anna’) (Dardano & Trifone, 1999). Additionally, plural and mass indefinite nouns may appear without a determiner, especially in generic or non-specific contexts.

The presence of multiple phonologically conditioned allomorphs, combined with the need to process both grammatical and phonological features in real time, makes article acquisition particularly challenging for L2 learners. Italian nouns ending in *-e* are especially problematic, as they often create ambiguity in gender assignment and article selection. This can result in errors such as (15):

(15) Bambara male speaker

*le	ellisse
l’	ellisse
the.DEF.F.SG	ellipse.F.SG
‘the ellipse’	

In this case, the learner overgeneralizes the plural feminine article to a singular feminine noun due to the ambiguous morphological cues.

Therefore, to better understand the acquisition challenges posed by such forms, it is essential to consider broader theoretical models that account for the grammaticalization of articles from lexical words into grammatical markers in typological and psycholinguistic terms. The grammaticalization of articles has been theorized through several models, each offering a distinct perspective on the evolution and functional development of nominal determination.

Table 1. below provides a comparative overview from a L2 acquisition perspective of three influential models: Greenberg's (1978) definite articles typological pathway, Heine's (1997) semantic continuum, and Givón's indefinite articles pathway (1981).

Table 1. Models of Articles' Grammaticalization

	Indefinite	Definite	
Stage	Givón (1981)	Greenberg (1978)	Heine (1997)
1	Expressing existence or quantity (quantification)	The demonstrative is used anaphorically; serves as the origin of the definite article	Articles originate from numerals indicating quantity or uniqueness (numeral stage)
2	Identifying specific entities (referentiality)	The demonstrative evolves into a definite article that marks identifiable referents	The numeral develops into a marker that identifies entities within discourse (presentative stage)
3	Non-referential, kind-level uses(genericity)	The article extends to specific but unidentified referents, entering the domain of indefinites. Usage is largely grammaticalized and governed by syntactic construction	The article signals specificity distinctions rather than quantity (specificity marker stage)
4	-	Referentiality is no longer relevant; the article becomes a mere gender marker or a nominal classifier, often accompanied by phonological reduction or affixation	The articles marks indefinite or non-identifiable referents, contributing to the grammar of indefiniteness (non-specific marker stage)
5	-	-	The article functions as an obligatory grammatical element, fully independent from its numeral origin (fully grammaticalized stage)

While these models differ in focus, they all trace the emergence of articles from functional precursors (e.g., demonstratives) to fully grammaticalized forms. Givón (1981) outlines a

developmental trajectory for indefinite articles, moving from quantification to referentiality and eventually to genericity. Greenberg (1978) conceptualizes the grammaticalization of definite articles as a cyclical process. However, as Kupisch and Polinsky (2021, p. 3) argue, a major limitation of Greenberg's model lies in the assumption that demonstratives evolve primarily into markers of gender rather than markers of definiteness. This critique is particularly relevant in the context of Italian, where definite articles serve a dual function: they not only encode definiteness but also carry morphosyntactic information, including gender and number agreement. It is also worth noting that in many languages, only one type of article, either definite or indefinite, is grammaticalized, which further complicates Greenberg's proposed linear pathway. Later, Heine (1997) proposes a semantic-referential continuum, moving from numerals to fully grammaticalized articles, and highlighting intermediate stages such as specificity. In the Italian context, empirical research on article acquisition has consistently identified a relatively stable developmental sequence that includes both definite and indefinite forms. Contrary to Heine's view, though, definite articles tend to emerge earlier than indefinite ones, both in terms of morphological simplicity and acquisition timing.

According to Chini (1995, p. 285), Italian L2 learners follow an implicational scale:

third-person anaphoric pronoun > definite article > indefinite article > attributive adjective > predicative adjective > past participle

Bernini (2010) similarly proposes that agreement features develop in the following order: third-person singular pronouns, definite articles, indefinite articles, attributive adjectives, predicative adjectives, and participial forms.

Pallotti (2005) further refines this perspective by noting that within definite forms, DEF.M.SG.*lo* and DEF.M.PL.*gli* tend to be acquired later than DEF.M.SG.*il* and DEF.F.SG.*la*, with the latter often emerging first and being overextended in early learner production. Indefinite articles, by contrast, appear later in the acquisition process.

These developmental patterns closely align with the frequency-based ranking of Italian determiners, as attested by the *Grande Dizionario della Lingua Italiana* (GDLI) (Note 1):

la (n=977573) > il (n=860845) > l' (n=495572) > le (n=485006) > un (n=621856) > i (n=374533) > una (n=421668) > lo (n=174431) > gli (n=203535) > un' (n=73718)

Data from the *Italian web corpus* (*itWaC*) (Note 2), a corpus composed of internet texts, shows similar frequency distributions:

il (n=30,293,839) > la (n=29,984,112) > un (n=16,402,719) > l' (n=15,997,284) > le (n=13,328,425) > i (n=13,320,663) > una (n=11,385,083) > gli (n=5,138,114) > lo (n=3,612,266) > un' (n=1,833,905) > uno (n=1,626,598)

The sources confirm the dominance of singular definite articles (*la, il, l'*), with DEF.F.SG.*la* and DEF.M.SG.*il* standing out as the most frequent forms. Notably, DEF.F.PL.*le* occurs more frequently than DEF.M.PL.*i* and DEF.M.PL.*gli* in the GDLI, highlighting the salience of feminine plural forms. In contrast, indefinite articles occur less frequently overall, with

INDF.M.SG.*un* more common than INDF.F.SG.*una*, and INDF.F.SG.*un'* being the rarest due to its restricted usage contexts.

This ordering supports a broader trend identified in both Italian and cross-linguistic studies: definite articles tend to emerge earlier in L2 acquisition because they are more frequent and less marked. Moreover, definite articles typically encode shared knowledge and accessible referents, a strategy that beginner learners often adopt. As De Marco (2005) observes, much like native-speaking children, adolescent L2 learners with low proficiency frequently assume shared discourse knowledge with the interlocutor, leading to overgeneralization of definite articles at the expense of indefinite ones.

1.3 Research Question

Based on the reviewed literature, this study addresses the following research question:

Can a developmental sequence in the acquisition of Italian articles in L2 be identified?

The central hypothesis posits that the acquisition of Italian articles by L2 learners follows a developmental trajectory influenced by frequency and markedness. Specifically, article forms that are more frequent in input and more consistently produced over time are expected to be acquired earlier and become more readily accessible in learners' interlanguage, irrespective of grammatical gender or number. From this perspective, definite articles are hypothesized to emerge earlier, as they are less structurally complex and more frequent, whereas indefinite articles are considered more marked and hence acquired later.

2. Method

To investigate the emergence and stabilization of Italian articles in L2 learner production over time, a quantitative time series analysis (Xu, 2023) for L2 research was conducted using the Python programming language. An analytical pipeline was implemented using the following libraries: *pandas* (McKinney, 2010) for data manipulation, *NumPy* (Harris et al., 2020) for numerical computing, and *statsmodels* (Seabold & Perktold, 2010) and *SciPy* (Virtanen et al., 2020) for statistical testing. Learner productions were timestamped, and a predefined list of articles specified the target forms to be tracked. The analysis focused on the output article, regardless of grammatical accuracy. Written production frequencies were computed by grouping the main data frame by date and output article, then aggregating occurrences. The resulting series was reshaped into a date-by-article matrix using *.unstack()*, with missing values filled using *fill_value=0*. To ensure uniform coverage, the resulting article counts were reindexed to include all articles in the predefined list. A total column was calculated by summing the article counts per date. The frequency of each article was then normalized by dividing by the daily total, resulting in a matrix of relative frequencies. Division-by-zero cases were handled by replacing *NaN* values with zero. To smooth daily fluctuations and reveal broader trends, a 7-day rolling average was computed using *.rolling('7D').mean()*. To identify acquisition sequences, the first stable emergence of each article was operationalized as a threshold of at least 5 occurrences within any 7-day window. For each article, a rolling sum over 7 days was computed from the raw article counts. If the threshold condition was met, the first date on which this occurred was recorded as the article's point of stabilization. Articles

were then sorted chronologically based on their first stable appearance to derive an acquisition sequence. Finally, statistical analyses were conducted to examine intergroup and temporal variation in article usage. These included one-way ANOVA (*stats.f_oneway*) to test whether the population means of all groups are statistically equivalent, and Tukey's Honest Significant Difference (HSD) (*pairwise_tukeyhsd*) (Seabold & Perktold, 2010), a post hoc procedure to conduct pairwise comparisons, evaluating whether the mean of each group significantly differs from that of every other group.

2.1 Participants

The participant pool consisted of 208 L2 learners of Italian (122 male; 86 female), aged between 11 and 29 years (M=15.07, SD=3.05), recruited from secondary schools in the Veneto region of northern Italy. Participants were recent migrants, with an average length of residence in Italy of 3.4 years (SD=4.5). The learners represented 35 distinct L1s, classified into four typological categories based on morphological structure: fusional (n=88), isolating (n=51), fusional-agglutinative hybrid (n=37), and agglutinative (n=32).

The classification of participants' L1s, as illustrated in Table 2, adheres to the parameters for article typology outlined by Dryer and Haspelmath (2013) in the *World Atlas of Language Structures* (WALS):

Table 2. Distribution of Participants' L1s According to Article System Type

Article Features	**n**	**Languages**
Full article system	13	Albanian, Bengali, French of Guinea, German, Jola, Mandinka, Moldovan, Nigerian Pidgin, Portuguese, Romanian, Spanish, Thai, Wolof
Indefinite-only	3	Chinese Mandarin, Persian/Farsi, Sinhalese
Definite-only	8	Bambara, Egyptian Arabic, Fula, Lebanese Arabic, Moroccan Arabic, Somali, Syrian Arabic, Tunisian Arabic
No article system	8	Bini, Bosnian, Dendi, Pashto, Russian, Shona, Tagalog, Ukrainian
Context-dependent system	3	Hindi, Nepali, Urdu

Context-dependent systems exhibit article-like behavior through demonstratives or quantifiers, rather than dedicated articles. All participants exhibited elementary proficiency in Italian, primarily due to limited access to formal instruction and restricted opportunities for language exposure. The linguistic profile of the sample reflects the broader sociolinguistic landscape of migration in contemporary Italy, where language isolation often persists despite nominal school enrollment.

2.2 Sampling Procedures

Data were collected during the 2024–2025 academic year within an Italian L2 laboratory implemented across eight secondary schools. In each school, the workshop spanned approximately three months and consisted of 16 hours of instructional time. Data collection was conducted using four complementary elicitation tasks to capture both controlled and spontaneous uses of Italian articles. The instruments included: (*i*) fill-in-the-blank tasks designed to elicit agreement in gender and number; (*ii*) binary-choice grammaticality judgment tasks, targeting article selection in minimal pairs; (*iii*) semi-structured written production tasks, including level-A2 peer interviews on daily life topics, informal emails to friends, and short narrative compositions; (*iv*) reading comprehension responses, focusing on article usage in contextually rich input. As Cornips & Poletto (2005) point out, elicitation contexts are to some extent artificial because participants are asked to produce a type of linguistic behavior that is a result of the elicitation design. In this study, tasks *(i)* and *(ii)* represent targeted elicited data, explicitly constructed to test hypotheses on determiner acquisition. These structured interventions provided a controlled yet ecologically valid setting for tracking learners' article production over time, allowing for systematic observation of emergent patterns across L1 backgrounds.

2.3 Research Design

The present study investigates the emergence and stabilization of Italian articles in the production of L2 learners over seventeen months. To this end, a quantitative, time-series approach was adopted to capture developmental trends and acquisition patterns across a corpus of 5341 written productions. The analysis centered on the article form produced by learners, regardless of grammatical accuracy, which served as the study's output variable. To operationalize the concept of acquisition, a stabilization criterion was introduced: an article was considered to have emerged stably when it appeared at least five times within any seven-day window. This threshold-based approach allowed for the identification of sustained rather than isolated usage. For each article, the first date meeting this criterion was recorded as its stabilization point. Determiner phrases were then sorted chronologically based on their first stable occurrence to reconstruct an empirically grounded sequence of acquisition.

3. Results

A total of 1995 errors and 3346 target-like productions were transcribed and tagged by date and article morphology. The acquisition scale is determined by calculating the first stable appearance that occurs every 7 days with a frequency of at least 5. The acquisition scale measures and captures the moment when an article becomes part of the student's frequent and

consistent production repertoire. It is based on the amount of use, i.e., the frequency, in the period 25/11/2023 - 23/4/2025, not on its intrinsic correctness. Learners may reach a stabilization point in the frequency of article usage within their interlanguage, despite the persistence of non-target-like realizations. The data reveal a clear developmental trajectory in the acquisition of Italian articles by L2 learners.

3.1 Statistics and Data Analysis

The resulting view of the relative frequency of articles over time shows how students' output is distributed among the various articles along the acquisition timeline (Figure 1).

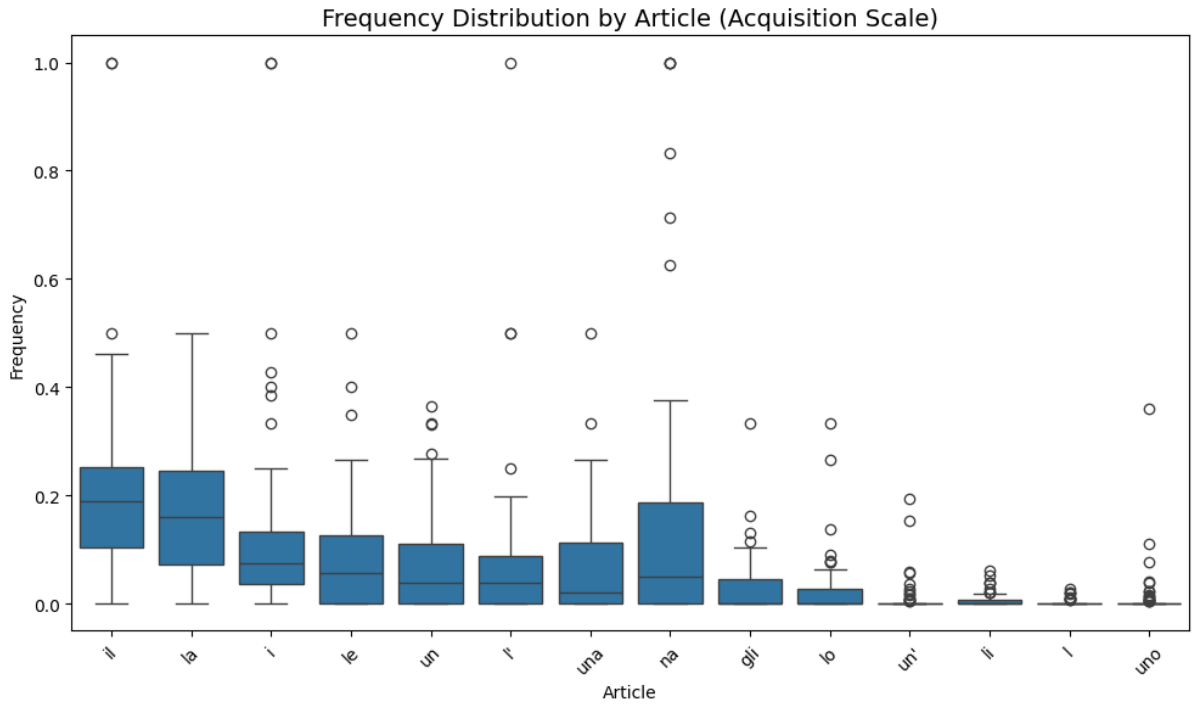

Figure 1. The Scale of Acquisition of Articles

The box plots display the distribution of relative frequencies for each item, ordered along the x-axis according to the acquisition scale determined before. Each box shows the median, quartiles, and outliers of relative frequency for a given article. The vertical position of the box and midline indicates the typical relative frequency with which that item is used.

The height, or interquartile range, and length of whiskers indicate variability in the frequency of use of that item. A narrower box indicates more consistent use. Comparing the box plots, it is observed that the articles acquired earlier, on the left, tend to have higher or more consistent relative frequencies than those acquired later, on the right. This might suggest that earlier acquired articles not only appear earlier but also become a more established and frequent part of the learners' output over time.

As exemplified, the articles' acquisition scale follows this order:

il > la > i > le > un > l' > una > na (omission) > gli > lo > un' > *li > *l > uno

A one-way ANOVA was conducted to test whether the relative frequencies of article production differed significantly across article types. The results indicate a statistically significant difference with a F-statistic of 38.49 ($p < 0.0001$). This means that learners do not produce all articles with the same frequency or rate: some articles are used significantly more or less often than others. This suggests that some articles were not only acquired earlier but also used more frequently than others, pointing to differential accessibility or salience in learner interlanguage.

To examine which article pairs differed significantly in terms of relative frequency, a Tukey HSD test was conducted following the significant one-way ANOVA result. The number of groups to compare with both correct forms and not is 26 (i.e., una l', a, al, e', el, gl, gli, i, il, in, l, l', la, le, li, ll, lo, na, nel, o, u, un, un', una, une, uno).

The formula yields the number of pairwise comparisons: $[n (n – 1)] / 2 = (26 \times 25) / 2 = 325$.

Therefore, out of a total of 325 pairwise comparisons, the analysis revealed that 155 were statistically noteworthy ($p < .05$, reject = True). Articles such as DEF.M.SG.*il*, DEF.F.SG.*la*, and INDF.M.SG.*un* showed significantly higher mean frequencies compared to less frequent and later-acquired forms like INDF.M.SG.*uno*, INDF.F.SG.*un'*, and DEF.M.PL.*gli*. For instance, the mean difference between DEF.SG.*l'* and DEF.F.SG.*la* was 0.0914, $p < .001$, indicating that the elided form behaves differently from the full form. Similarly, the mean difference between DEF.M.SG.*il* and INDF.F.SG.*una* is –0.1368, $p < .001$, reflecting a robust contrast between the articles: the negative sign of the mean difference suggests a relative prominence of DEF.M.SG.*il* compared to INDF.F.SG.*una*. Conversely, no significant differences emerged between INDF.F.SG.*una* and DEF.SG.*l'* (meandiff = -0.0111, $p = 1.0$), or between DEF.M.PL.*gli* and INDF.F.SG.*un'* (meandiff = -0.0208, $p = 0.9987$), hinting at similarities in their distributional patterns. For example, INDF.M.SG.*un* and INDF.F.SG.*una* exhibited a non-significant (reject = False) mean difference (meandiff = -0.0122, $p = 1.0$), indicating that these two forms do not differ significantly despite their gender distinction.

These post hoc results confirm the acquisition sequence observed in the time-series analysis: articles that stabilize earlier tend to be produced more frequently. Results highlight concurrently expected contrasts between high-frequency and low-frequency articles, as well as subtle patterns, such as the distinct behavior of elided versus non-elided definite forms. This reinforces the hypothesis that production frequency is closely tied to the stabilization of grammatical forms in interlanguage development.

4. Discussion

Based on the developmental stage proposed and supported by the present statistical evidence, a descriptive analysis discussion of learner errors is presented.

During the earliest developmental phase, learners predominantly produce omission errors (n = 597), as illustrated in (16). This type of error occupies a central position on the acquisition scale (na), as it occurs throughout the initial stages, peaking in December 2023.

(16) Syrian female speaker

*con	mia	familia	
con	la	mia	famiglia
with.COM	the.DEF.F.SG	my.1SG.POSS	family.F.SG

'with my family'

This may suggest an avoidance strategy, or negative transfer from L1 -ART. This first (or zero) phase of omission would correspond to what Chini and Ferraris (2003, cited in Schmiderer & Hinger, 2023, p. 44) have defined as the 'pragmatic' or 'phonological phase', in which almost no determinative articles are used, and morphological variation does not follow recognizable rules (Note 3). Similar patterns are observed in child language, where articles are optionally omitted, as well as in adult speakers with specific types of brain damage, such as agrammatic Broca's aphasia (Menn & Obler, 1990, among others). Comparable omissions also occur in so-called 'special registers' used by adults, i.e., newspaper headlines or colloquial speech (De Lange, 2008).

In the subsequent stages of the acquisition scale, namely the lexical and proto-morphological phases, a notable increase in definite and indefinite article production is observed. During these phases, learners begin to experiment with article use, often retrieving forms from memory as lexical chunks. In spite of that, article morphology remains underdeveloped, resulting in highly heterogeneous errors, particularly in gender and number agreement. This stage is characterized by partial rule awareness, as reflected in the data; errors reach their highest frequency at this point, indicating that learners show emerging sensitivity to article presence but are unable to apply morphosyntactic rules consistently.

The results reveal a systematic overuse of the feminine DEF.F.SG.*la* (n=206) (17) as well as the DEF.M.SG.*il* (n=233) (18), consistent with findings from previous studies (Chini, 1995; Pallotti, 2005; Nitti, 2023), throughout all the 35 L1.

(17) Chinese male speaker

*la	calcolo
il	calcolo
the.DEF.M.SG	calculation.M.SG

'the calculation'

(18) Tagalog female speaker

*il	Groenlandia
la	Groenlandia
the.DEF.F.SG	Greenland.F.SG

'the Greenland'

A frequency-based analysis reveals that DEF.M.SG.*il* and DEF.F.SG.*la* are exploited by all 35 L1 with an average use of 7.7%. This overgeneralization also extends to plural contexts, with learners incorrectly applying singular forms in both feminine (19) and masculine (20) plural environments.

(19) Bengali female speaker

*la vacanze
le vacanze
the.DEF.F.PL holidays.F.PL
'the holidays'

(20) Farsi female speaker

*il patate
le patate
the.DEF.F.PL potatoes.F.PL
'the potatoes'

Importantly, from the analyzed frequencies, this tendency is quite well-distributed across all L1 backgrounds. Learners with L1s such as Albanian, Bosnian, Egyptian Arabic, French of Guinea, Fula, Moroccan Arabic, Nigerian Pidgin, and Urdu show a marked tendency to overuse DEF.F.SG.*la*; in contrast, speakers of Bengali, Bini, Farsi, German, Mandinka, Romanian, Shona, Sinhalese, Spanish, Syrian, Tunisian Arabic, and Wolof more frequently overuse DEF.M.SG.*il*. No strong preference emerges for the remaining L1 groups.

Furthermore, the error distribution reveals that contexts requiring DEF.M.SG.*il* are more susceptible to substitution errors (n=430) than those requiring DEF.F.SG.*la* (n=271), suggesting greater instability in the processing and retrieval of the masculine singular article.

Nevertheless, the early emergence of a form in interlanguage does not necessarily indicate higher accuracy. To determine which article exhibits greater mastery, an accuracy rate was calculated. This measure considered the number of correct productions relative to the total number of obligatory contexts in which the article was required, including correct and erroneous realizations. The accuracy rate was computed only for items that appeared at least once in the learner data:

$$\text{accuracy\_rate} = \text{correct\_productions} / \text{total\_required} \ (\text{if total\_required} > 0)$$

Results are presented in Table 3.

Table 3. Accuracy Rates for Definite Article *il* and *la*

Index	Correct Production	Incorrect Where Required	Omission Where Required	Total	Accuracy Rate	%
il	883	220	210	1313	0.672506	67.25
la	769	156	115	1040	0.739423	73.94

As shown, DEF.F.SG.*la* demonstrates a higher accuracy rate (73.94%) and thus greater stability in learners' output compared to DEF.M.SG.*il* (67.25%). When DEF.F.SG.*la* is required, students are more likely to produce it correctly than DEF.M.SG.*il* in its obligatory context.

This finding aligns with previous research. Indeed, earlier studies (Berretta, 1990; Chini, 1995) have documented an overextension of the morpheme *-a* in the early stages of article acquisition. This phenomenon is often linked to the perception of *-a* as a prototypically Italian vowel. DEF.F.SG.*la* forms a consonant-vowel (CV) structure, which corresponds to what phonological theory identifies as the universally unmarked syllable type (Hume, 2004). The overuse of DEF.F.SG.*la* in Italian L2 does not appear to result from unsystematic errors, but rather from developmentally predictable processes grounded in perceptual salience and articulatory simplicity. As Jakobson (1968) argued, the low central vowel [a], acoustically salient and phonetically open, is among the first acquired in both L2 and L1 phonological development (see also Pizzuto & Caselli, 1992, on L1 acquisition). This may account for the privileged status of DEF.F.SG.la in early interlanguage grammars: its morphophonological transparency probably makes it easier to access and retrieve than masculine forms, which involve morphophonological conditioning and attention to gender agreement.

Nonetheless, the frequent use of DEF.M.SG.*il* challenges a purely markedness-based explanation. If syllable structure were the sole factor, one might expect DEF.M.SG.*lo*, also CV, and arguably less frequent, to appear more prominently (cf. Repetti, 2020). Inversely, DEF.M.SG.*il*, a VC syllable more phonologically marked than DEF.F.SG.*la* and DEF.M.SG.*lo*, remains highly frequent in productions, which may explain its early emergence in learner grammars. Interestingly, learners show a clear preference for DEF.M.SG.*il* over DEF.M.PL.*i* (n=48), suggesting that number distinctions are more marked in learner production than gender distinctions.

The simultaneous early presence of both DEF.F.SG.*la* and DEF.M.SG.*il* supports the hypothesis that frequency effects and markedness together influence articles' acquisition. As Greenberg (1966) observed, within any opposition of linguistic items, the more frequent member tends to function as the unmarked one. In this vein, these highly frequent articles are more likely to be acquired early, functioning as default or unmarked forms in the learner's developing grammar (Ellis, 2002).

Curiously, the overuse of the DEF.M.SG.*il* and DEF.F.SG.*la* mirrors the tendency to assign grammatical gender within complex determiner phrases, and particularly those containing possessive adjectives, based on the biological or referential gender of the possessor rather than the grammatical gender of the noun itself. Regardless of whether L1 learners initially rely on prioritizing natural gender over syntactic agreement (21, 22):

(21) Bengali male speaker

*il	mio	madrelingua
la	mia	madrelingua

the.DEF.F.SG my.1SG.POSS mother tongue.F.SG

'my mother tongue'

(22) Farsi female speaker

*la mia corsi

i miei corsi

the.DEF.M.PL my.1SG.POSS courses.M.PL

'my courses'

This cognitive economy suggests that universal processing strategies are primary. What is more, DEF.F.SG.*la* and DEF.M.SG.*il* exhibit morphosyntactic (e.g., DEF.M.SG.*lo*; DEF.M.PL.*gli*) or phonologically conditioned variants (e.g., DEF.M.SG.*l'* and DEF.F.SG.*l'* before vowels). This invariance likely contributes to its cognitive and processing advantages in L2 acquisition: learners face fewer morphophonological constraints and decision points when using the default forms of masculine and feminine articles (Dressler, 1985), thereby reducing processing costs in lexical retrieval (Levelt, 1989). Such findings are consistent with the broader cross-linguistic observation that all learners acquire unmarked forms as a necessary developmental stage before acquiring marked forms (White, 1987, p. 266).

It is therefore unsurprising that the patterns observed for these two forms align closely with their high token frequency in large-scale Italian corpora, such as GDLI and itWaC.

In the same way, also the occurrence of DEF.F.PL.*le* (n=186), DEF.M.PL.*i* (n=200), and the intermediate error form **li* (n=46), corresponding to (C)V syllables attested in 14 out of 35 languages, can be explained in articulatory terms: they are phonetically more accessible, less marked, even for children, than forms such as DEF.M.PL.*gli* (CCV) or DEF.M.SG.*lo* (CV) (De Marco, 2005, p. 77). DEF.M.PL.*i* is overused by Bosnian, Chinese, Sinhalese, and Wolof speakers as a strategy to avoid the DEF.M.PL.*gli*, accounting for a total of 110 non-target-like forms, whereas the overuse of DEF.M.PL.*gli* over DEF.M.PL.*i* occurs in only 27 non-target-like forms. Furthermore, DEF.M.PL.*i* is overapplied relative to DEF.M.SG.*il* (n=31) and DEF.F.PL.*le* (n=39). In contrast, Bambara, Moroccan Arabic, and Ukrainian students tend to prefer the DEF.F.PL.*le*, which occurs in 186 non-target-like forms, showing the highest overuse: 53 forms over the DEF.F.SG.*la* and 43 over DEF.M.PL.*i*.

Results of the accuracy rates for DEF.M.PL.*i* and DEF.F.PL.*le* are presented in Table 4.

Table 4. Accuracy Rates for Definite Article *i* and *le*

Index	Correct Productions	Incorrect Where Required	Omission Where Required	Total Required	Accuracy Rate	%
i	241	217	69	458	0.526201	52.62
le	246	179	34	425	0.578824	57.88

The data show that mastery of the correct use of the DEF.F.PL.*le* is at a slightly more advanced or robust level (57.88%) than that of the DEF.M.PL.*i* (52.62%), confirming that the vowel /e/ in DEF.F.PL.*le* is more accessible than /i/ for some students, including Arabic speakers. As Giacalone Ramat (2003) already pointed out, over-extensions are the most significant concern of the DEF.F.SG.*la* and DEF.F.PL.*le*. As acknowledged by Jakobson (1968, p. 30), /i/ and /e/ can function as free variants of the same narrow phoneme, the first variant being stronger and more distinct, more distant from the wide phoneme, and the other being weaker and less distinct.

What is more, because Italian allows adjectives to occupy prenominal and postnominal NP positions, the form of the determiner cannot be specified until the major constituents of the phrase are ordered. It is only at this point that the phonological context relevant for determiner selection (either the onset of the noun or an adjective) can be ascertained (Miozzo & Caramazza, 1999, p. 920). Therefore, in noun classes ending in *-e*, the DEF.F.PL.*le* is frequently used as an overgeneralization error, like in (23):

(23) Romanian male speaker

*le	fonte
la	fonte
the.DEF.F.SG	source.F.SG
‘the source’	

It is not a coincidence that the hybrid form **li* (n=46) is exploited by Albanian, Bengali, Chinese, Egyptian Arabic, Farsi, Fula, Jola, Mandinka, Moldovan, Moroccan Arabic, Tunisian Arabic, Ukrainian, Urdu, and Wolof. /i/ may be exploited as an epenthesis after the /l/ as less marked than /o/, according to the markedness scale based on Chomsky & Halle (1968). Accordingly, it may emerge as an interlinguistic phonological compromise: it appears in DEF.M.PL.*gli* (n=8) context as an avoidance of the complex /ʎ/ sound (absent in many L1s, e.g., Chinese and Arabic). It appears also in all the other contexts, DEF.M.SG.*il* (n=17) (24), DEF.F.PL.*le* (n=9), DEF.SG.*l’* (n=7), DEF.F.SG.*la* (n=3), and DEF.M.PL.*i* (n=2).

(24) Wolof male speaker

*li	mondo
il	mondo
the.DEF.M.SG	world.M.SG
‘the world’	

It is also plausible that **li* reflects an overgeneralization of the frequent form *i* + *le*, with /i/ taking precedence over /e/, as predicted by Jakobson’s (1968) vowel hierarchy.

In the subsequent developmental stage, INDF.M.SG.*un* emerges as one of the most stable and frequent forms. Its semantic overlap with the numeral ‘one’ makes it cognitively salient and

thus more easily acquired than other indefinite articles. As noted by Valentini (1990, cited in Gudmundson, 2012, p. 35), INDF.M.SG.*un* is often preferred over its morphological variants (see example 25), although among the incorrect outputs, INDF.F.SG.*una* (26) emerges as the second most frequently produced form (n=97 non-taget-like forms; n=288 targe-like forms) after INDF.M.SG.*un* (n=109 non-taget-like forms; n=415 targe-like forms), likely due to the phonological factors outlined above. In contrast, the delayed acquisition of INDF.F.SG.*un'* and INDF.M.SG.*uno*, after January 2024, may stem from their lower input frequency and more restricted syntactic distribution.

(25) French male speaker from Guinea

*un	sirena
una	sirena
a.INDF.F.SG	mermaid.F.SG
'a mermaid'	

(26) Moroccan female speaker

*una	ragazzo
un	ragazzo
a.INDF.M.SG	boy.M.SG
'a boy'	

In a later morphosyntactic stage, learners demonstrate increased stability in article usage. Nonetheless, sporadic errors persist, possibly due to idiomaticity or residual L1 transfer. For example, although in standard Italian the definite article is omitted before singular family member nouns, the example in (27) nevertheless displays correct agreement between the article, possessive, and noun.

(27) Shona male speaker

*il	mio	fratello
Ø	mio	fratello
Ø	my.1SG.POSS	brother.M.SG
'my brother'		

This suggests that the learner has attained a higher level of control over morphosyntactic structures.

5. Conclusion

This study investigated whether Italian L2 learners follow a developmental trajectory in the acquisition of articles and whether unmarked and high-frequency forms emerge earlier. The results support Greenberg (1966) and Eckman's MDH (1985), revealing that unmarked,

high-frequency, and phonologically simple articles such as DEF.M.SG.*il* and DEF.F.SG.*la* are acquired earlier, followed by DEF.F.PL.*le* and DEF.M.PL.*i*. Among these, feminine forms, which follow a CV structure, show greater stability than their masculine counterparts. The vowels [a], [e], and [i], considered acoustically salient, are acquired before more marked vowels such as /o/ in DEF.M.SG.*lo*, in line with Jakobson's (1968) markedness hierarchy. Among indefinite articles, INDF.M.SG.*un* is preferred and acquired earlier than INDF.M.SG.*uno*, INDF.F.SG.*un'*, and INDF.F.SG.*una* due to its higher perceptual salience and lower morphophonological complexity. In contrast, intermediate marked forms (l', lo, gli, uno, un') show delayed stabilization, have higher error rates, and are produced with significantly lower frequency (five to ten times less) than default forms, likely because of their morphophonological constraints.

Methodologically, the study relies solely on written productions, which may differ from spoken language patterns. Consequently, future research could extend the analysis to spoken production to capture modality-specific acquisition trajectories and provide a more comprehensive account of article emergence. Moreover, unequal productivity among learners may bias frequency estimates: less productive individuals contribute little data, while highly productive learners may disproportionately shape group-level trends. To address the issue of uneven learner productivity, subsequent studies might employ mixed-effects models that account for inter-individual variability in production. Finally, the use of rolling averages, while highlighting broad tendencies, may obscure short-term fluctuations. Future studies could employ mixed-effects models to account for inter-individual variability and dynamic modeling approaches (e.g., generalized additive or state-space models) to preserve short-term patterns while capturing overall developmental trends.

Nonetheless, these findings contribute to broader discussions on morphosyntactic transfer and L2 development by highlighting interactions between phonological markedness, morphological complexity (e.g., gender-number concord), and frequency effects. Students' choices are influenced not only by morphosyntactic accuracy but also by phonological simplicity and input salience. As such, learners may produce morphologically plausible forms, yet non-native-like productions. The degree to which learners from different L1 backgrounds navigate these stages remains an open question. Future research should examine whether speakers of -ART L1s (e.g., Russian) follow different developmental paths compared to learners whose L1s include articles (e.g., Spanish), further elucidating the interplay between L1 transfer and markedness in L2 acquisition.

Educationally, the outcomes suggest structuring instruction from the most frequent and perceptually salient forms to more marked, infrequent forms, with explicit attention to discriminating article allomorphs to support stable acquisition.

Acknowledgements

Sincere thanks are due to the teachers who welcomed the author into their classrooms and supported the implementation of the Italian language laboratory.

References

Andrews, E. (1990). *Markedness Theory. The Union of Asymmetry and Semiosis in Language*. Durham: Duke University Press Durham and London.

Bernini, G. (2010). Acquisizione dell'italiano come L2. In R. Simone (Ed.), *Enciclopedia dell'italiano* (pp. 139-140). Roma: Istituto dell'Enciclopedia Italiana G. Treccani.

Berretta, M. (1990). Morfologia in italiano lingua seconda. In E. Banfi, & P. Codin (Eds.), *Storia dell'italiano e forme dell'italianizzazione: atti del XXIII 131 Congresso internazionale di studi: Trento - Rovereto 18-20 maggio 1989* (pp. 181-201). Roma: Bulzoni.

Bybee, J. (2010). Markedness: Iconicity, Economy and Frequency. In J. J. Song (Ed.), *Handbook of Linguistic Typology* (pp. 131-147). Oxford: Oxford University Press.

Chini, M. (1995). *Genere grammaticale e acquisizione: aspetti della morfologia nominale in italiano L2*. Milano: FrancoAngeli.

Chomsky, N., & Halle, M. (1968). *The Sound Pattern of English*. New York: Harper & Row.

Cornips, L. & Poletto, C. (2005). On standardising syntactic elicitation techniques (part 1). *Lingua*, *115*, 939-957.

Dal Pozzo, L. (2022). Definiteness and Indefiniteness. A Comparative Perspective on Finnish and Italian. *Lea*, *11*, 355-372.

Dardano, M., & Trifone, P. (1999). *Grammatica italiana: con nozioni di linguistica*. Bologna: Zanichelli.

De Lange, J. (2008). Article Omission in Headlines and Child Language: A Processing Approach. *Doctoral dissertation*. LOT Series, Utrecht University.

Dressler, W. (1985). On the Definite Austrian and Italian Articles. In E. Gussman (Ed.), *Phono-Morphology: Studies in the Interaction of Phonology and Morphology* (pp. 36-47). Lublin: Catholic University.

Dryer, M. S., & Haspelmath, M. (Eds.). (2013). WALS Online (Version 2020.4) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.13950591

Eckman, F. R. (1985). Some Theoretical and Pedagogical Implications of the Markedness Differential Hypothesis. *Studies in Second Language Acquisition*, *7*(3), 289-307. Retrieved from http://www.jstor.org/stable/44488563

Ellis, N. C. (2002). Frequency Effects in Language Processing: A Review with Implications for Theories of Implicit and Explicit Language Acquisition. *Studies in Second Language Acquisition*, *24*(2), 143-188.

Giacalone Ramat, A. (2003). *Verso l'italiano. Percorsi e strategie di acquisizione*. Roma: Carocci.

Giacalone Ramat, A. (2008). Typological Universal and Second Language Acquisition. In S. Scalise, E. Magni, & A. Bisetto (Eds.), *Universals of Language Today* (pp. 1-20). Berlin: Springer.

Giusti, G. (1997). The Categorial Status of Determiners. In L. Haegeman (Ed.), *The New Comparative Syntax* (pp. 95-123). London: Longman.

Givón, T. (1981). On the Development of the Numeral 'One' as an Indefinite Marker. *Folia Linguistica Historica*, *2*(1), 35-53.

Greenberg, J. H. (1966). *Language Universals, with Special Reference to Feature Hierarchies*. The Hague, The Netherlands: Mouton.

Greenberg, J. H. (1978). How does a language acquire gender markers? In J. H. Greenberg, C. A. Ferguson, & E. Moravcsik (Eds.), *Universals of human language* (pp. 47-82). Stanford, CA: Stanford University Press.

Gudmundson, A. (2012). *L'accordo nell'italiano parlato da apprendenti universitari svedesi: Uno studio sull'acquisizione del numero e del genere in una prospettiva funzionalista*. Stockholm: Department of French, Italian and Classical Languages, Stockholm University, Forskningsrapporter/Cahiers de la recherche.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357-362. https://doi.org/10.1038/s41586-020-2649-2

Heine, B. (1997). *Cognitive Foundations of Grammar*. Oxford: Oxford University Press.

Hume, E. (2004). Markedness: A Predictability-Based Approach. *Proceedings of the Annual Meeting, Berkeley Linguistics Society*, *13*, 182-198.

Jakobson, R. (1968). *Child Language, Aphasia and Phonological Universals*. The Hague: Mouton.

Kupisch T., & Polinsky, M. (2021). Language history on fast forward: Innovations in heritage languages and diachronic change. *Bilingualism: Language and Cognition*, 1-12. https://doi.org/10.1017/S1366728921000997

Levelt, W. J. M. (1989). *Speaking: From intention to articulation.* Cambridge, MA: MIT Press.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, Austin, 56-61.

Menn, L., & Obler, L. (Eds.). (1990). *Agrammatic aphasia: A cross-language narrative sourcebook*. Philadelphia: John Benjamins.

Miozzo, M., & Caramazza, A. (1999). The Selection of Determiners in Noun Phrase Production. *Journal of Experimental Psychology: Learning, memory, and cognition*, *25*(4), 907-922.

Nitti, P. (2023). L'articolo determinativo fra acquisizione e insegnamento. *Kwartalnik Neofilologiczny, 70*(4), 527-545.

Odlin, T. (1989). *Language Transfer: Cross-Linguistic Influence in Language Learning*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139524537

Pallotti, G. (2005). Le ricadute didattiche delle ricerche sull'interlingua. In E. Jafrancesco (Ed.), *L'acquisizione dell'italiano L2 da parte di immigrati adulti* (pp. 43-59). Atene & Roma Edilingua.

Pizzuto, E., & Caselli, M. C. (1992). The Acquisition of Italian Morphology: Implications for Models of Language Development. *Journal of Child Language*, *19*(3), 491-557.

Repetti, L. (2020). The masculine singular definite article in Italian: The role of the syllable. *Italian Journal of Linguistics*, *32*(2), 209-232.

Schmiderer, K., & Hinger, B. (2023). L'Interlingua Produttiva e Ricettiva di Studenti di Italiano LS in un Contesto di Scuola Secondaria Austriaca. *Italiano LinguaDue*, *15*(2), 43-64. https://doi.org/10.54103/2037-3597/21938

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. *Proceedings of the 9th Python in Science Conference*, *57*(61), 92-96. https://doi.org/10.25080/Majora-92bf1922-011

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., … SciPy 1.0 Contributors. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods*, *17*(3), 261-272. https://doi.org/10.1038/s41592-019-0686-2

Ward, G., & Birner, B. (1995). Definiteness and the English existential. *Language*, *71*(4), 722-742.

White, L. (1987). Markedness and Second Language Acquisition: The Question of Transfer. *Studies in Second Language Acquisition*, *9*(3), 261-285. Retrieved from http://www.jstor.org/stable/44487416

Xu, D. (2023). Time Series Analysis as an Emerging Method for Researching L2 Affective Variables. *Heliyon*, *9*(6), e16931. https://doi.org/10.1016/j.heliyon.2023.e16931

Glossary

Ø = omission; 1 = first person; ART = article; COM = comitative; DEF = definite; F = feminine; INDF = indefinite; M = masculine; M = mean; meandiff = mean difference; n = total number; p = probability value; PL = plural; SD = standard deviation; SG = singular; POSS = possessive.

Notes

Note 1. https://www.gdli.it/elenco-forme-per-frequenza

Note 2. https://www.sketchengine.eu/itwac-italian-corpus/

Note 3. The second phase is the lexical one, the third is the proto-morphological, and the fourth is morphosyntax.

Copyrights