CODIT. A new resource for the study of Italian from a diachronic perspective: Design and applications in the morphological field

CODIT. Une nouvelle ressource pour l'étude de l'italien dans une perspective diachronique: applications dans le domaine morphologique

M. Silvia Micheli

Electronic version
URL: https://journals.openedition.org/corpus/7306
DOI: 10.4000/corpus.7306
ISSN: 1765-3126

Publisher
Bases ; corpus et langage - UMR 6039

Electronic reference
M. Silvia Micheli, "CODIT. A new resource for the study of Italian from a diachronic perspective: Design and applications in the morphological field", Corpus [Online], 23 | 2022, Online since 02 March 2022, connection on 05 March 2022. URL: http://journals.openedition.org/corpus/7306 ; DOI: https://doi.org/10.4000/corpus.7306

This text was automatically generated on 5 March 2022.

© Tous droits réservés
1. Introduction

This paper presents a new resource for the study of Italian from a diachronic perspective, i.e., CODIT (Corpus Diacronico dell’Italico ‘Diachronic corpus of Italian’), namely a balanced corpus containing about 29 million occurrences. In the first part of the article, we describe the corpus structure and discuss its features compared to another diachronic corpus of Italian (i.e., MIDIA, Morfologia dell’italiano in diacronia ‘Italian morphology in diachrony’). In the second part of the article, we provide the analysis of two case studies to show how CODIT can contribute to the study of morphological phenomena from a diachronic perspective. Both cases deal with evaluative morphology, which so far has been mostly investigated from a synchronic point of view: in particular, we will outline the diachrony of the prefix supra- (e.g., sopravvalutare ‘to overestimate’) and the suffix -oide (e.g., sferoide ‘spheroid’). Both affixes allow us to illustrate the dynamism that characterizes the domain of evaluative morphology, where evaluative values (such as diminution, augmentation, pejoration) undergo a cyclic renewal to preserve their expressiveness (see Grandi 2011).

The paper is structured as follows. Section 2 provides a detailed description of the corpus design. Section 3 offers a diachronic analysis of supra- and -oide in Italian, based on both qualitative and quantitative data extracted from CODIT. Section 4 focuses on the potential and the limits of the corpus and outlines future developments.
2. Corpus design

CODIT is a balanced diachronic corpus of written Italian of about 29 million occurrences, hosted by the Czech National Corpus website, where it can be queried through the Kontext interface. It represents a revised version of the CODIT_com corpus, which has been collected for the diachronic study of Italian compounding provided in Micheli (2020). Although it has initially been collected with the goal of allowing morphological investigations from a diachronic perspective, it can be considered a general-purpose corpus. At present, the corpus has not been lemmatized nor POS-tagged, but both steps are planned for the near future, as we will discuss in more details in Section 4.

The texts collected within CODIT cover a period ranging from the earliest attestations of the Italian language to 1947. It is structured into five subcorpora depending on the chronological period. The periodization follows that adopted for the MIDIA corpus (see Iacobini, De Rosa & Schirato 2017): it is based on important linguistic and social facts of Italian history. Specifically, the five chronological periods are the following:

1) 13th century-1375: this subcorpus covers a period ranging between the earliest attestations of Italian and Boccaccio’s death.
2) 1376-1532: this subcorpus encompasses Humanism and Renaissance. It ends in 1532 with the publication of the third edition of the Orlando furioso by Ludovico Ariosto.
3) 1533-1691: this subcorpus encompasses literary Mannerism and Baroque. It ends in 1691 with the publication of the third edition of the Vocabolario by the Accademia della Crusca.
4) 1692-1840: this subcorpus encompasses the Enlightenment and Romantic period. It ends in 1840 with the publication of the final edition of the Promessi Sposi by Alessandro Manzoni.
5) 1841-1947: this subcorpus represents a period ranging from the Risorgimento to the end of the Second World War. It ends in 1947 with the publication of the Italian Constitution.

Each subcorpus collects texts belonging to six genres, i.e., essays, literary prose, poetry, letters, scientific texts, theatre. The only exception is the first subcorpus, which does not include scientific texts. As we will show in Section 3.2.2, the presence of texts of different genres is crucial to shed light on semantic changes occurring in word-formation as well as on the relationship between ordinary language and scientific terminologies.

As far as the corpus size is concerned (see Table 1), each subcorpus includes around 6 million tokens, except for the first subcorpus, which consists of 4.5 million tokens, due to the difficulty of collecting electronic form texts. The normalization of the frequency data extracted from CODIT is advised, given the slightly different size of the subcorpora. The structure of the corpus is illustrated in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>essays</td>
<td>1,545,178</td>
<td>575,328</td>
<td>1,914,096</td>
<td>1,057,300</td>
<td>1,167,999</td>
</tr>
<tr>
<td>letters</td>
<td>28,023</td>
<td>1,120,431</td>
<td>982,128</td>
<td>1,582,788</td>
<td>1,312,679</td>
</tr>
</tbody>
</table>

Table 1. CODIT: structure and size
The CODIT’s design is heavily indebted to MIDIA. In particular, the main debt to MIDIA deals with the identification of the five chronological periods, which has been adopted without modification. On the other hand, the choice of text genres shows a slight difference: differently from CODIT, MIDIA includes a subcorpus containing legal and administrative texts. Moreover, the selection of texts that make up the corpus does not coincide with that carried out for MIDIA.

Notably, the main difference between the two resources concerns size, since CODIT consists of 29 million occurrences, while MIDIA contains 7.5 million occurrences. Small size represents a significant limitation to analyzing linguistic phenomena that are not particularly frequent in texts, such as many morphological phenomena (both inflectional and derivational). The main goal of CODIT is to provide a large amount of data belonging to different domains, in order to capture a wide range of phenomena and linguistic variation.

Moreover, it should be noted that MIDIA comprises portions of texts containing 8,000 tokens, while, within CODIT, each text has been included in its entirety.

Texts have been collected in .txt format and cleaned from spelling inconsistencies. Each text is associated with a set of metadata, i.e., doc ID, title, author, year (or century, when the dating is unclear), chronological period, textual genre, variety (this latter parameter applies only to the first two subcorpora, corresponding to the period before the normalization of literary Italian proposed by Pietro Bembo and based on the Florentine variety). As an example, Table 2 illustrates metadata associated to Storia d’Italia by Francesco Guicciardini.

Table 2. Metadata associated to Storia d’Italia by Francesco Guicciardini

<table>
<thead>
<tr>
<th>p.id</th>
<th>805</th>
<th>doc.period</th>
<th>1533-1691</th>
</tr>
</thead>
<tbody>
<tr>
<td>doc.id_file</td>
<td>3_ESP_GUICCIARDINI_STORIA D’ITALIA</td>
<td>doc.genre</td>
<td>ESPOSITIVI</td>
</tr>
<tr>
<td>doc.author</td>
<td>Francesco Guicciardini</td>
<td>doc.title</td>
<td>Storia d’Italia</td>
</tr>
<tr>
<td>doc.year</td>
<td>1561</td>
<td>doc.variety</td>
<td></td>
</tr>
</tbody>
</table>

Thanks to the Kontext interface, the corpus can be queried using both simple and advanced queries (through regular expressions); restricted queries based on metadata are also possible.

As already mentioned, texts are still raw, non-annotated, thus the corpus can be searched only for forms (not for POS-tags or lemmas). Annotation represents the main further step of this project (see Section 4).
In the following Sections, we provide the diachronic analysis of two evaluative affixes based on data extracted from CODIT. These case studies allow us to illustrate possible uses of the corpus in the study of word-formation from a diachronic perspective.

3. Case studies

3.1. Theoretical issues

According to Grandi (2002: 31-34), the semantic functions performed by evaluative morphemes can be classified into two classes, i.e., ‘quantitative’ evaluation and ‘qualitative’ evaluation. The former encompasses the values ‘diminution’ and ‘augmentation’, which imply a descriptive evaluation based on objective properties. The latter includes a more significant number of values, namely intensification, endearment, prototypicality, attenuation/approximation, contempt, which express an evaluation based on a subjective and personal judgement by the speaker. Although in the last decades evaluative morphology has been extensively investigated (see Grandi & Körtvélyessy 2015), not all semantic functions performed by evaluative morphemes have been equally addressed, especially within the field of qualitative evaluation. Focusing on Italian, while intensification has been recently investigated (see Grandi 2017, for an overview), less attention has been paid to approximation (see Masini & Micheli 2020). Except for Napoli’s (2012, 2017) studies on the intensive prefix stra- (e.g., strapieno ‘chock-full’), intensification and approximation have so far only been investigated from the synchronic perspective. Notably, the diachronic perspective is crucial to capture the dynamism that characterizes the domain of evaluative morphology. With the aim of enriching the literature on the emergence (and the decay) of evaluative meanings, we outline the diachrony of the prefix sopra- and the suffix -oide throughout the history of Italian.

3.2. Basic facts on sopra- and -oide

In Contemporary Italian, the prefix sopra- (also occurring as sovra-) is homophonous to the preposition sopra ‘above, over, on’ (from Latin sūpra ‘id.’), which expresses a locative value and, more rarely, the meaning ‘about X, concerning X’. According to Iacobini (2004: 132, 139, 149, 154), sopra- can convey the following values:

(1)


b. time ‘after’: sopraggiungere ‘suddenly arrive’, sopravvivere ‘survive’

c. quantity ‘larger amount, excess amount’: sovrannumero ‘supernumerary’, sopstattassa ‘additional fee’

d. repetition ‘again’ (very rare): soprastampare ‘overprint’

As shown by these examples, sopra-/sovra- can combine with adjectives (especially relational adjectives), nouns and verbs. Some of these words are already attested in the earliest stages of the language and show a certain degree of lexicalization (i.e., sopressedere ‘overlook’). Intensification is not mentioned among the values illustrated in (1); however, Iacobini (2004: 385) mentions that when sopra- combines with qualifying adjectives like in sopraffino it conveys the value ‘very’. Notably, Grandi & Montermini (2005) mentioned sopra-/sovra- among those prefixes used to express an evaluative meaning (which now is no longer productive according to the authors) in the previous
stages of the language. This point deserves further attention and will be addressed in our analysis. In the last decades, sopra-/sohra- seems scarcely productive: sopra- words are not attested within the two principal repositories of neologisms, i.e., the Treccani Dictionary of Neologisms and the repository provided by the Osservatorio neologico della lingua italiana (ONLI). As far as -oide is concerned, it comes from Lat. -oide(m), in its turn arising from Ancient Greek -oïdēs (cf. eidos ‘shape’). According to Rainer (2004: 263), -oide primarily is an adjectival suffix expressing similarity (e.g., antropoide ‘human-like in appearance’, antropo- ‘human’ -OIDE). In scientific terminology, it often derives also nouns from nouns (e.g., metalloide ‘metalloid’, metal-OIDE). Both in ordinary language and specialized domains, -oide conveys the value ‘X is similar to Y’ (where Y is the base of the derived word). Wandruszka (2004: 400) includes -oide words in the class of ‘disposition adjectives’, namely adjectives referring to human referents based on an inclination/tendency that they show to the entity conveyed by the base (e.g., sinistroide ‘leftist’, left-OIDE). As noted by Ricca (2004: 434), -oide shares some features with other simulative suffixes, such as -esco (e.g., spagnolesco ‘Spanish-like’) and -eggiante (e.g., orientaleggiante ‘oriental-like’). According to Merlini Barbaresi (2004: 449-450), -oide originally occurred in scientific terms (mainly in medicine, geometry, anthropology languages). The spread in ordinary language is more recent and entails the emergence of a pejorative value since what is imperfect (i.e., not precisely coincident with X) is considered as worse than X. The expression of evaluative meanings by affixes originally occurring in scientific terminology has been already observed in Italian by Masini & Micheli (2020) in the cases of simil- (e.g., temperature simil-estive ‘summery-like temperatures’) but also of para- ‘id.’, quasi- ‘almost’, and semi- ‘half’, all expressing a set of values related to approximation. Our analysis will contribute to shed light on the path towards evaluation followed by affixes originally expressing a categorizing/classificatory function.

3.3. Methodology

Our analysis is based on data extracted from CODIT. The corpus has been queried through the Kontext interface for forms that start with sopra- and forms that end with oide(i/a/o). As far as sopra- is concerned, all variants (namely, sopra-, supra- and sohra-) have been considered. A manual check is needed to exclude false positives (e.g., soprano). Finally, the selected forms have been manually lemmatized. Data have been analyzed according to the following parameters. From the formal standpoint, we have considered the lexical category of the base as well as the lexical category of the complex word. Semantically, we have analyzed the values conveyed by sopra- and -oide; a qualitative analysis of each context has been crucial for the interpretation of the value conveyed by the two affixes.

3.4. Results

3.4.1. Sopra-

As illustrated by Table 3, we have started off our analysis by calculating both type and token frequencies of sopra- within the five subcorpora of CODIT.
Data reported in Table 3 highlights that *sopra*- words are attested in all five periods of CODIT. As far as type frequency is concerned, the first period contains the highest number of *sopra*- words; after a slight decrease corresponding to the subcorpora 2-3-4, the number of types increases in the last period. A similar path can be observed in token frequency.

Our investigation continues with a more in-depth analysis of types, according to the lexical categories of the bases and the semantics conveyed by *sopra*-. The results are illustrated in Table 4.

Table 4. *Sopra*- words (type frequency, raw data): lexical categories of the bases and semantics

<table>
<thead>
<tr>
<th>Lexical category of the base</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADJ</td>
<td>10</td>
<td>4</td>
<td>8</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>ADV</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>NOUN</td>
<td>14</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td>35</td>
</tr>
<tr>
<td>V</td>
<td>31</td>
<td>33</td>
<td>28</td>
<td>31</td>
<td>35</td>
</tr>
<tr>
<td>PRON</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Semantics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOCATIVE</td>
<td>29</td>
<td>34</td>
<td>38</td>
<td>33</td>
<td>50</td>
</tr>
<tr>
<td>QUANTITY</td>
<td>12</td>
<td>15</td>
<td>12</td>
<td>17</td>
<td>20</td>
</tr>
<tr>
<td>TIME</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>INTENSIFICATION</td>
<td>12</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

The analysis revealed that *sopra*- mainly combines with verbs (e.g., *sopravincere* ‘to prevail’, SOPRA+win) and nouns (e.g., *sopravveste* ‘clothes worn over other garments’, SOPRA+clothes), more rarely with adjectives (e.g., *soprabello* ‘very good’, SOPRA+good). In three cases, *sopra*- combines with an adverb (i.e., *soprapiù* ‘surplus’, SOPRA+plus) and a pronoun (i.e., *sopraciò* ‘responsible for administrative functions’, SOPRA+anything, and soprattutto ‘above all’, SOPRA+all). Moreover, while the presence of verbal bases is stable over the centuries, the combination of *sopra*- and nouns is more frequent in the last period.

As far as the lexical categories of the outputs are concerned, the combination of *sopra*- with a base generally does not change the lexical category of the base (e.g., SOPRA+grande_{ADJ} ‘big’ > sopragranda_{ADJ} ‘very big’). However, we have found the following
cases containing sopra- and a noun (or a pronoun in the case of soprattutto) that function as adverbs: 1) sopraramano (SOPRA+hand) ‘openly’ (e.g., parlare sopramano ‘to speak openly’) and ‘from top to bottom’ (e.g., colpire sopramano ‘to hit from top to bottom’); 2) soprammercato (SOPRA+trade) ‘what is more’; 3) soprammodo (SOPRA+manner) ‘extremely’; 4) soprattutto (SOPRA+all) ‘above all, especially’. All these words always occur as adverbs, except for sopraramano, which also functions as a noun referring to a stitching type. Moreover, we can also mention soprapiiù and sopracciò (see above), where sopra- combines with (respectively) an adverb and a pronoun to create nouns.

Turning to the semantics of sopra-, we have found that sopra- most frequently conveys a locative meaning (i.e., ‘above, over, on’ as in soprastruttura ‘superstructure’, SOPRA+structure) that sometimes can be figured (e.g., sopradivino ‘more than heavenly’, SOPRA+heavenly’). Throughout the centuries, the value QUANTITY (i.e., ‘excess amount, additional’; e.g., sopraperizia ‘additional expert opinion’) becomes more frequent. When sopra- expresses a temporal meaning, it conveys the idea that a given event has now taken place or takes place suddenly, as in the cases of soprarriunto ‘just arrived’ or sopraprendersere ‘catch suddenly’, exemplified in (3).

<table>
<thead>
<tr>
<th>(3) a.</th>
<th>Un di coloro si staccò dalla brigata, s’accostò al soprarriunto e gli domandò se veniva da Milano.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>one of those REFLECT separate.3.SG.PFV of the brigade REFLECT approach.3.SG.PFV to the SOPRA-arrived and he REFLECT ask.3.SG.PFV if come.3.SG.PFV from Milan</td>
</tr>
<tr>
<td></td>
<td>‘One of those separated from the brigade, approached one who had just arrived and asked him if he was coming from Milan’. (Alessandro Manzoni, I Promessi Sposi, ed. 1840)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>b.</th>
<th>giunto a l’uscio de la casa paterna, ode la voce de parenti, onde sente sopraprendersi da una certa letizia [...].</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>arrived.REFLEX.PTCP.PST to the door of the house of his father hear.3.SG.PRS the voice of relatives for which feel.3.SG.PRS</td>
</tr>
<tr>
<td></td>
<td>‘When he arrives in front of the door of his father’s house, he hears the voice of his relatives and is suddenly caught by a certain joy’. (Pietro Aretino, La Talanta)</td>
</tr>
</tbody>
</table>

In (3a), soprarriunto refers to someone who had arrived at that moment, unexpectedly; similarly, in (3b), sopraprendersere indicates that the character is suddenly caught by a feeling of joy as soon as he arrives at the front door. The value ‘suddenly’ can also be
attributed to words like sopraggiungere (SOPRA+arrived) or sopravvenire (SOPRA+come), both meaning ‘to arrive unexpectedly’, in which, on the contrary, Iacobini (2004; see Section 3.2) identified the value ‘after’. Among the sopra-words attested in CODIT, the value ‘after’ only occurs in sopravvivere ‘survive, continue to live after a lethal event or after the death of other people’ which comes from Latin supravivere ‘id.’. Notably, sopra- can also express an evaluative meaning, namely intensification, as in the cases of sopragridare and sopranobile.

<table>
<thead>
<tr>
<th>(4) a.</th>
<th>Tosto che parton l’accoglienza, prima che l’ombrone sopragrida, ciascuna s’affatica.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>once that leave.3.PL.PRS the welcome, before that the shadow sopra-shout.3.SG.PRS each one toil.3.SG.PRS</td>
</tr>
</tbody>
</table>

‘Immediately after separating from that friendly welcome, even before moving a step away from each other, each host of souls strives to shout out’. (Dante Alighieri, Commedia)

b. Di capo di queste tre giornate, si trova la sopranobile città di Quinsai [...]. E conterò di sua nobiltà, però che è la più nobile città del mondo e la migliore. |

‘At the end of these three days, the excellent city of Quinsai can be found. And I will tell you about its nobility since it is the noblest and best city in the world’. (Marco Polo, Milione)

In (4a), sopragridare indicates that the souls in purgatory described by Dante shout out loud so as to overpower others’ voices: in this case, sopra- conveys the idea of excess, referring to an action that has gone beyond a specific limit. In (4b), the Chinese city of Quinsai (today’s Hangzhou) is described by Marco Polo as sopranobile, namely excellent, very noble, and a little further on as la più nobile ‘the noblest’. In this case, sopra-
attributes to the base a meaning comparable to that of the superlative -issimo. As illustrated in Table 3, the intensifying function of *sopra-* is well attested in the first period while becomes marginal in the following centuries. It should be noted that already in Old Italian, the use of *sopra-* as intensifying prefix is mainly restricted to adjectives (i.e., 11 adjective, seven verbs, one noun) and mainly occurs in literary texts (especially poetries): in most cases, these words are hapaxes that do not occur in other text genres. The only exception is the adjective sopraffino ‘excellent, masterly’ (SOPRA-fine) which is attested from the third period and shows a significant frequency in the last subcorpus (i.e., 19 occ.). Currently, it shows a certain degree of lexicalization, also due to the low frequency of the base fino ‘elegant, fine’.

In Table 5, we have analyzed the relationship between the semantics of *sopra-* and the lexical categories of the bases.

Table 5. Semantics of *sopra-* and lexical categories of the bases (raw data)

<table>
<thead>
<tr>
<th>Lexical categories of the bases</th>
<th>LOCATIVE</th>
<th>QUANTITY</th>
<th>TIME</th>
<th>INTENSIFICATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adj</td>
<td>10</td>
<td>3</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>Adv</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Noun</td>
<td>38</td>
<td>23</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Pronoun</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Verb</td>
<td>45</td>
<td>10</td>
<td>7</td>
<td>7</td>
</tr>
</tbody>
</table>

As illustrated in Table 5, the LOCATIVE value of *sopra-* is well attested in nouns and verbs, while the QUANTITY value mainly occurs in nouns, especially nouns referring to quantifiable things (e.g., soprassoldo ‘additional compensation’). As already mentioned, the evaluative function of *sopra-* is more frequent when it combines with qualifying gradable adjectives (e.g., bello ‘good, nice’, acuto ‘smart’, mirabile ‘marvellous’, piacevole ‘pleasant’) and verbs (e.g., gioire ‘rejoice’, vincere ‘win’, gridare ‘shout’). Finally, we can observe that the temporal value is almost exclusively attested when *sopra-* combines with verbs.

To summarize, based on data extracted from CODIT, we have found that starting from Old Italian *sopra-* conveys different values when it combines with verbs, adjectives, nouns (and, more rarely, with pronouns and adverbs). The primary function of *sopra-* is to express a locative indication ‘above, over’, according to the meaning of the free preposition/adverb *sopra*. We have also noted that *sopra-* can convey the idea of excess, which can be interpreted as ‘excess amount of X, additional X’ (e.g., sopralavoro ‘excessive work’ soprattassa ‘additional fee’) or as ‘very X’ (e.g., sopragioioso ‘very joyful’) depending on the lexical category and the semantics of the base. In particular, our analysis highlighted that the evaluative, intensifying interpretation can be primarily found when *sopra-* combines with qualifying (gradable) adjectives, while the QUANTITY value is often attested with nouns referring to quantifiable entities. Finally, we have found that *sopra-* can also convey a temporal value and indicate that a given event has just occurred.

The decreasing number of words where *sopra-* expresses intensification can be related to the history of another originally locative prefix that acquired an intensifying function and became one of the most frequent morphological means to express this...
value, namely stra- (e.g., strafamoso ‘very popular’, STRA+popular). As pointed out by Napoli (2012), stra- (from Latin adverb/preposition extra ‘outside, beyond’) shows an intensifying value already in Old Italian, mostly in combination with qualifying adjectives. It shows significant similarities with sopra-, as demonstrated by the fact that, in some cases, they attach to the same bases (e.g., strabundanza and soprabbondanza ‘superabundance’). Thus, they can be considered in competition in the expression of intensification in adjectives, at least in the most ancient stages of the language. A further in-depth comparison between sopra- and stra- (as well as between sopra- and other intensifying elements such as super-) can be carried out based on data extracted from CODIT.

### 3.4.2. -oide

Turning to -oide, we have calculated the type and token frequencies of -oide words in each subcorpus of CODIT.

**Table 6. -oide: type and token frequencies (normalized data in brackets)**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td>45</td>
</tr>
<tr>
<td>0 (1)</td>
<td>(0.9)</td>
<td>(9.7)</td>
<td>(40.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Token</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>1</td>
<td>99</td>
<td>178</td>
<td></td>
</tr>
<tr>
<td>0 (2)</td>
<td>(0.9)</td>
<td>(96)</td>
<td>(159)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As illustrated in Table 6, -oide words are not attested in the first subcorpus. The first attestations of complex words containing -oide as suffix have been found in the second and third subcorpora (corresponding, respectively, to the periods 1376-1532 and 1533-1691). Specifically, the oldest type we have found is romboide ‘parallelogram that is neither rhombus nor rectangle’, attested in the ancient feminine form romboida (used as an adjective referring to the noun figura ‘figure’) and the masculine noun romboido. The geometrical term romboide represents a learned loan from Ancient Greek ῥομβοειδής (see Latin rhomboides) and was introduced by Euclid to refer to a type of parallelograms which are similar in shape to rhombuses. Within the subcorpus 3, we only found the adjective concoide (a loan from Ancient Greek κογχοειδής where κόγχη ‘shell’) used in physics to indicate a type of fracture with a curved surface.

Starting from the fourth period, -oide begins to occur more frequently in complex words: we have focused on the last two subcorpora due to increased data availability. Subcorpus 4 contains 10 types, six of which belong to the geometry vocabulary, namely cicloide ‘cycloid’, conoide ‘conoid’, ellissoide ‘ellipsoid’, paraboloidoide ‘paraboloid’, romboide ‘rhomboid’, sferoide ‘spheroid’. The four exceptions are still part of scientific terminology, e.g., ovoide ‘ovoidal’ (egg-OIDE) and aracnoide ‘arachnoid mater’ (aracn-’spider’) are used in medical vocabulary to indicate body parts or organs showing a shape similar to that conveyed by the base. Within subcorpus 5, we have found 45 types containing -oide, mostly attested in scientific texts. In particular, 29 types only occur in scientific texts, four types occur both in scientific texts and essays or letters, while 14 types only occur in essays.
From a morphological point of view, we have noted that -oide mostly combines with autonomous words (e.g., \textit{paraboloide} ‘paraboloid’, \textit{parabola} ‘id.’ +OIDE), but in 16 cases, it occurs with combining forms (e.g., \textit{aracnoide} ‘arachnoid mater’ where \textit{ara}- ‘spider’ is a combining form of classical origin). Moreover, the bases are mostly morphologically simplex words (e.g., \textit{negroide} ‘negroid’, \textit{negro} ‘black’ +OIDE), but some cases where the bases are derived words are attested (e.g., \textit{criminaloide} ‘potential criminal’ where \textit{criminale} ‘criminal’ is derived from \textit{crimine} ‘crime’ through the suffix -\textit{ale}). In some rare cases, -oide does not add to a derived base; rather, it replaces its suffix, as in \textit{epilettoide} ‘epileptoid, referring to clinical manifestations reminiscent of those of epilepsy’, made up of \textit{epilett(ico)} ‘epileptic’ +OIDE).

Finally, we have analyzed the lexical categories of the bases with which -oide combines and its semantics.

Table 7. -oide: lexical categories of the bases and semantics in the last two subcorpora (raw data)

<table>
<thead>
<tr>
<th>Lexical categories of the bases</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADJ</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>NOUN</td>
<td>10</td>
<td>36</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Semantics</th>
<th>SHAPE</th>
<th>SIMILARITY</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>16</td>
</tr>
</tbody>
</table>

As illustrated in Table 7, within the fourth subcorpus, -oide always combines with nouns and refers to ‘something similar to X in shape’, where X is the entity conveyed by the base.

In the last subcorpus, -oide mainly occurs with nouns, while in nine cases, it combines with adjectives (especially relational adjectives), all indicating human beings (e.g., \textit{borghesoide} ‘who poses as bourgeois’). From a semantic point of view, -oide can indicate similarity in shape\textsuperscript{10} (e.g., \textit{ovoide} ‘ovoideal, egg-like in shape’) –namely the original value conveyed in Ancient Greek and attested in the learned loans– or express a broader notion of similarity, based on various parameters. In particular, in scientific texts, -oide is used with a categorizing function to refer to entities or materials which are similar but not identical to a given entity or material (e.g., \textit{granitoide} ‘referring to igneous rocks with characteristics similar to those of granite’ or \textit{alcaloide} ‘organic substance with properties similar to those of alkalis’). Within political or social essays, -oide still refers to similarity based on concrete/physical features in anthropological terms related to the classification of types of \textit{homo sapiens}, defined based on the shape of the skull and other craniometric and anthropometric features (e.g. \textit{negroide} ‘negroid’ or \textit{australoid} ‘australoid’). These cases are crucial in that they allow both interpretation (i.e., similarity strictly based on shape and similarity based on other concrete features). Finally, -oide can refer to someone that behaves like a member of a given (social) group but without being part of it, such as in \textit{socialistoide} ‘sympathizer of socialism, pseudo-socialist’ or \textit{mattoide} ‘weirdo, psycho’ (crazy+OIDE), exemplified in (5).
(5) a. I governi di molti paesi d'Europa [...] difficilmente possono respingere tutte queste e le analoghe aspirazioni dei socialisti e socialistoidi. (Gaetano Mosca, Elementi di scienza politica)

‘The governments of many European countries [...] can hardly reject all these and similar aspirations of the socialists and pseudo-socialists’.

b. che direbbero gli insigni psichiatri, così facili a dispensare la patente di matto o di mattoide, se qualcuno affermasse invece che i mattoidi sono loro? (Olindo Guerrini, Brani di vita)

‘What would the illustrious psychiatrists, that easily dispense the license of madman or psycho, say if someone said instead that they are psycho?’.

In both examples, the derived word occurs together with the corresponding non-derived word to highlight that the speaker is referring to both the prototypical member of a given category (e.g., the true madman) and the peripheral elements that are similar but not coincident with the prototype (e.g., the psycho). In these derived words, the value conveyed by -oide is evaluative in nature in that it reflects the speaker’s judgment, which can show an approximating or a pejorative value. This approximating value is well-attested in the last decades, as illustrated in the following examples, extracted from the itTenTen16 corpus, i.e., a web corpus of Contemporary Italian.

(6) a. Dalle nostre parti quello slang non è mai approdato, nonostante il massiccio utilizzo di terminologia inglesoide.

‘That slang has never landed where we come from, despite the massive use of English-like terminology’.

b. [... con i suoi soldi il ministero dell'Interno finanzierà un osservatorio carfagnoide sulla minoranza degli omosessuali e le altre a rischio di discriminazione.

‘[...] with his money, the Ministry of the Interior will fund a Carfagna-style observatory on homosexuals and other minorities at risk of discrimination’.

Both examples in (6) show that -oide is currently used by speakers to indicate something (or someone) that resembles –but does not coincide with– a given category. Interestingly, the base can also be represented by a proper noun, as in (6b), where carfagnoide refers to an observatory in line with the other initiatives against gender-based violence promoted by the politician Mara Carfagna (former Italian minister for Equal Opportunities). In both examples, -oide also conveys a pejorative nuance, making the speaker’s tone polemic.

To sum up, we have found that -oide firstly occurs in a small number of learned loans belonging to the geometry vocabulary where it conveyed the original Ancient Greek value related to shape. Starting from the period represented by the fourth subcorpus, the frequency of -oide starts to increase: it occurs in scientific terminologies where it combines only with nouns and always expresses the value ‘similar in shape’. Data extracted from the most recent period represented within CODIT show a significant increase in terms of frequency; moreover, we have found that -oide also occurs in essays and letters, where it combines with both nouns and adjectives. From a semantic point of view, it is used (even if not frequently) also to express the evaluative function of approximation (and sometimes pejoration). The usage of -oide in the first part of the 20th century is crucial to understand the current behaviour of -oide in ordinary language, characterized by a loosening of the semantic and morphological constraints on the selection of the base. As already mentioned, the emergence of an evaluative meaning in affixes initially used in scientific terminologies has been already noted by Masini & Micheli (2020) for Italian. This analysis confirms that affixes initially used with a categorizing function can develop an evaluative value and that the emergence of the approximating function represents a recent change.
4. Concluding remarks

This article has outlined the main features of CODIT by focusing on its structure in subcorpora, based on the chronological period and the text genre. This structure proved to be crucial to understand the semantic changes that have affected two affixes, i.e., the prefix *sopra-* and the suffix *-oide*. The corpus size has allowed us to extract a good amount of data for the quantitative analysis, while metadata provided for each text has made it possible to contextualize each occurrence of *-oide* or *sopra-* words. Finally, the lack of annotation and lemmatization has made the data preparation more time-consuming but did not prevent to carry on this investigation. Nonetheless, further steps will be devoted to lemmatization and POS annotation of the corpus.

BIBLIOGRAPHY


NOTES

1. We have chosen to compare CODIT with MIDIA since the latter represents the only other balanced diachronic corpus which covers the entire history of the Italian language and includes texts of different genres. However, as noted by an anonymous reviewer, other resources have been used in previous diachronic studies on Italian, such as LIZ 4.0 (Stoppelli & Picchi 2001) and DiaCORIS (Onelli et al., 2006).


4. The morphological status of words such as soprammobile (SOPRA+piece of furniture) has been discussed within the literature devoted to the boundaries between compounding and derivation. According to Scalise (1994: 136), they are exocentric compounds, in that the head of the word does not coincide with one of the two constituents (i.e., a sopramobile IS not A [type of] mobile). On the other hand, Montermini (2008: 139-48) ascribes sopra- to a set of bisyllabic prefixes (among which also contro- ‘counter/against’, sotto- ‘sub/under’) which show non-prototypical features. In this study, we have considered these words as prefixed and included them in our analysis.

5. According to Iacobini (2004: 157), in sovrastampare ‘overprint, print over another print’, sopra-show an iterative function. However, we suggest that also locative interpretation is possible here.


7. The ONLI repository can be accessed at the following URL: https://www.iliesi.cnr.it/ONLI/ (accessed: 29/09/2021).

8. Normalized data in brackets.

9. However, it should be noted that romboide is attested within the corpus of Old Italian OVI (Opera Vocabolario Italiano), which is larger than the first subcorpus of CODIT.

10. These words have been classified separately since the relationship between the derived word and the base rests on a specific kind of similarity related to a concrete feature, namely the shape.
ABSTRACTS

This paper presents a new diachronic corpus of the Italian language, i.e., CODIT (Corpus Diacronico dell’Italiano ‘Diachronic corpus of Italian’). After a description of the corpus design, the second part of the article is devoted to show possible applications of the corpus in the morphological field, specifically in the domain of evaluative morphology. We highlight the corpus’ strengths and limitations through the diachronic analysis of two affixes, i.e., sopra- and -oide. The study contributes to shed light on the dynamism of evaluative morphology by focusing on the loss of the intensification function suffered by sopra- and the emergence of an approximating value that currently affects -oide.

Cet article présente un nouveau corpus diachronique de la langue italienne : CODIT (Corpus Diacronico dell’Italiano « Corpus diachronique de l’italien »). Après une description du corpus, la deuxième partie de l’article est consacrée à montrer les applications du corpus dans le domaine morphologique, en particulier dans le domaine de la morphologie évaluative. Nous soulignons les forces et les limites du corpus à travers l’analyse diachronique de deux affixes, sopra- ‘sur-’ et -oide ‘-oïd’. L’étude contribue à éclairer le dynamisme de la morphologie évaluative en s’intéressant à la perte de la fonction d’intensification subie par sopra- et à l’émergence d’une valeur d’approximation qui affecte actuellement -oide.

INDEX

Mots-clés: morphologie évaluative, corpus diachronique, formation de mots, italien
Keywords: evaluative morphology, diachronic corpora, word-formation, Italian

AUTHOR

M. SILVIA MICHELI

Università degli studi di Milano – Bicocca