8

Studi Linguistici Pisani

Strutture linguistiche e dati empirici in diacronia e sincronia

a cura di

Giovanna Marotta e Francesca Strik Lievers

PISA
UNIVERSITY
PRESS

Strutture linguistiche e dati empirici in diacronia e sincronia / a cura di Giovanna Marotta e Francesca Strik Lievers. - Pisa : Pisa university press, 2017. - (Studi linguistici pisani ; 8)

417.7 (22.)
I. Marotta, Giovanna II.Strik Lievers, Francesca 1. Linguistica storica 2. Semantica e sintassi 3. Apprendimento 4. Neuropsicologia

CIP a cura del Sistema bibliotecario dell'Università di Pisa

Società con socio unico Università di Pisa
Capitale Sociale € 20.000,00 i.v. - Partita IVA 02047370503
Sede legale: Lungarno Pacinotti 43/44 - 56126 Pisa
Tel. +39 050 2212056 - Fax +39 050 2212945
e-mail: press@unipi.it
www.pisauniversitypress.it

ISBN: 978-88-6741-789-6

Indice

III. Acquisizione e perdita

Story of a Construction: statistical and distributional analysis of the development of the Italian Gerundival Periphrases

Irene Amato e Alessandro Lenci

1. Introduction

Diachronic linguistics in particular represents an ideal ground for corpus-based analyses, which allow linguistic change to be seen as a sequence of gradual changes in distributional patterns of usage (Bybee 2010: 118). This study aims to examine the diachronic development of the semantic structure of the Italian *Gerundival Periphrases*. This group of formally related constructions consists of the *Progressive Construction*, which is expressed by a sequence of the verb *stare* 'to stay' followed by a gerund, and the *Continuous Construction*, which can be further distinguished in two variants, one with *andare* 'to go', the other with *venire* 'to come', both followed by a gerund (Bertinetto 1989-90). The methodological foundation of the present research is provided by two theoretical approaches firmly rooted in a usage-based perspective: *Construction Grammar* (Goldberg 1995; Hoffman and Trousdale 2013), which regards grammar as emerging from patterns of usages, and *Distributional Semantics* (Lenci 2008), which assumes that relevant aspects of meaning are related to textual co-occurrences and therefore are continuous and gradient. Based on the information extracted from the *Google Ngram Corpus* (Michel *et al.* 2011), this research has observed which verbs have been combined with these three constructions (Progressive, Continuous$_{andare}$, Continuous$_{venire}$) during eight centuries of documentation, in order to infer some conclusions about the construction's semantics and productivity, investigating how it changes in time.

Various statistical and distributional semantic analyses are presented, all rooted in a usage-based view of grammar and meaning, to tackle three main issues. Firstly, we have measured the syntactic productivity of these argument structure constructions, checking their evolution in time. Secondly, we have analyzed the clusters of verbs that occur with each construction. This has led to recognize semantic cores that highlight interesting similarities and differences among the Gerundival Constructions. Finally, we have created distributional semantic spaces for each construction within various temporal slices that

show the evolution of their productivity, a key concept to understand how constructions arise, grow and change. In comparison with the previous studies on the topic, this one is based on a larger amount of data, which cover a very long period thanks to the contribution of two diachronic corpora[1], provides a distributional representation of the constructional semantics and is particularly focused on productivity.

2. Italian Gerundival Constructions

With *construction*, we refer to "any linguistic pattern, where some aspects of its form or function is not strictly predictable from its component parts or from other constructions, or fully predictable with sufficient frequency" (Goldberg 2006: 5). The linguistic objects of this study are the Italian constructions in which a gerund forms a morphosyntactic and semantic unity with a partially semantically weakened verb, whose original meaning is related either to localization or to movement. These structures can be referred to as the Progressive Construction, characterized by the verb *stare* 'to stay', and the two Continuous Constructions, formed by the verbs *andare* 'to go' or *venire* 'to come' (Bertinetto 1989-90: 29-47):

(1) a. *La ballerina* sta danzando.
'The dancer *is dancing*.'
b. *La folla* va aumentando.
'The crowd *is increasing*.'
c. *La nave* si viene avvicinando.
'The ship *is approaching*.'

These periphrases can be considered as a group not only from the syntactic point of view, but also from the semantic one. As far as the aspect is concerned, they all convey an imperfective meaning, as the incompatibility with the telic adverbial *in x time* proves (Bertinetto 1997: 172-173):

(2) **Paolo* stava/andava/veniva dipingendo *la parete* <u>*in due ore*</u>.
'*Paolo *was painting* the wall <u>within two hours</u>.'

More precisely, in combination with telic verbs, these periphrases assume a meaning of incrementative progression, while with non telic predicates they simply express plain imperfectivity (Squartini 1990: 209):

[1] Here, it has been possible to present only the results from *Google Ngram Corpus*. However, the queries have been performed on *MIDIA corpus* too (cf. note 2).

(3) a. *La temperatura* sta aumentando *di giorno in giorno.*
'The temperature *is increasing* day after day.'
b. *Paolo* sta dormendo *da ore.*
'Paolo *has been sleeping* for hours.'

Moreover, the gerundival periphrases are interchangeable in contexts expressing gradual completion (Bertinetto 1997: 168):

(4) a. *Questi guanti* si stanno/vanno/vengono *a poco a poco* infeltrendo.
'Little by little, these gloves *are getting matted.*'
b. *Filippo* stava/andava/veniva assomigliando *sempre più a suo zio.*
'Filippo *resembled* more and more his uncle.'

On the other hand, several features distinguish the three constructions. Firstly, the semiauxiliary *stare* introduces a punctual focalization with an incidental interpretation. Through this monofocalization, the construction expresses the progressive aspect, becoming "an imperfective aspectual marker denoting a situation as on-going at a given contextually relevant time point" (Squartini 1998: 79). On the other hand, both Continuous Constructions imply a durative perspective and expresses the continuous aspect through plurifocalization (Bertinetto 1989-90: 46-47). They require a reference time-space, along which the action expressed is distributed in every instant (Squartini 1998: 237-243).

(5) a. *Istante dopo istante, Paolo* andava/veniva/*stava annotando *le sue impressioni.*
'Minute by minute, Paolo *was writing* his impressions *down.*'
b. *Quando squillò il telefono, Paolo* stava/?andava riordinando *gli appunti.*
'When the telephone rang, Paolo *was reorganizing* his notes.'

Moreover, due to the motion meaning of *andare* and *venire,* the Continuous Constructions describe the process of approaching (and not necessarily reaching) a goal, resulting in an actional valence with decreased telicity, as the incompatibility with the telic adverbial *in x time* shows (Bertinetto 1997: 172-173). However, it is also true that the gerund undergoes a partial increase of telicity thanks to this meaning of motion towards a goal: the periphrasis transforms activities into incrementative predicates (cf. 6a) and preserves the telicity of those achievements that have been made durative (cf. 6b) (id.: 165-166):

(6) a. *Filippo* andava scrutando *l'orizzonte in cerca di navi corsare.*
'Filippo *was observing* the horizon, looking for pirate ships.'

b. *Il deposito* andava esplodendo.
'The depository *was exploding*.'

Moving to the analysis of the difference between the two variants of the Continuous Construction. *Andare* + *gerund* conveys a modular dimension of variation, as the frequent recurrence of adverbs indicates. Modal expressions of graduality (e.g. *con sempre maggiore x* 'with more and more x'), temporal expressions of graduality (e.g. *giorno dopo giorno* 'day by day'), manner adverbs (e.g. *insistentemente* 'insistently') often appear with this periphrasis. It is remarkable how *andare* + *gerund* instances gain in acceptability through graduality and modality expressions (adverbs, but also inherently intensified verbs). On the other hand, *venire* + *gerund* simply underlines the focus on the goal or on the recipient of the represented event (Bertinetto 1989-90: 42-44). As it can be seen in the following examples, contextual modality, graduality and iterativity do not suffice to trigger this construction:

(7) a. *La nave* andava/*veniva *inesorabilmente* affondando.
'The ship *was* inexorably *sinking*.'
b. *Sara* andava/??veniva mangiando *sempre di più*.
'Sara *was eating* more and more.'
c. *Maria* andava/??veniva scribacchiando *sul bordo del quaderno*.
'Maria *was scribbling* on the edge of the notebook.'

The *venire* variant shows a more marked and better preserved deictic meaning and acquires a less grammaticalized structure, which is goal-oriented. Due to this, it is possible to find the Continuous$_{venire}$ Construction only in durative and telic contexts. Orientation to the telos is the sole requirement, which can be expressed also through the presence of a recipient of the action (ibid.):

(8) *Luigi* si veniva/?andava guadagnando *gli spiccioli facendo il facchino*.
'Luigi *was earning* pocket money working as a porter.'

3. Data and methods

The diachronic development of the three constructions has been investigated with data extracted from the *Google Ngram Corpus* (Michel *et al.* 2011).[2] This corpus is freely available for eight different languages, among which Italian. It contains 305,763 digitalized Italian volumes, with a total of 40,288,810,817 Italian words. The text distribution is not balanced: for the XVI century, only

[2] http://books.google.com/ngrams/

few publications per year are available. However, the information about the total number of tokens per year allows to compare data from different time frames. Information about the textual typology has not been considered, but it is plausible that it gets more varied with time, containing more literary and standardized texts for previous phases and more different production for the contemporary age. The instances of the three constructions have been extracted automatically, without checking them manually, by retrieving only adjacent sequences of *semiauxiliary* + *gerund*. After the query, 305,081 occurrences of these constructions from 1550 to 2009 have been obtained: 131,193 instances of *andare* + *gerund* (43%), 126,123 of *stare* + *gerund* (41.34%), 47,765 of *venire* + *gerund* (15.66%).[3]

4. Statistical analysis: results

Each plot in Figures (1)-(4) contains three lines, which represent the three constructions: in red *andare* + *gerund*, in blue *venire* + *gerund*, in green *stare* + *gerund*. The diachronic development of the following indices of productivity[4] has been detected. Token frequency counts the total number of occurrences of a construction. Since it shows how frequently a linguistic structure is used, it is an index of its cognitive entrenchment and correlates with the possibility of analogical extensions. Type frequency represents the total number of distinct verb types that appear in a construction each year. The higher it is, the greater the productivity is. The curve of growth plots for each year the cumulative number of verb types that have appeared with a construction up to that year. The Type/Token Ratio is computed as the ratio between the number of verb types appearing with a construction and the token frequency of that construction. It indicates the degree of lexical variety of the construction's schematic[5] slot (in the present case, the verb) and typically correlates with its productivity.

---

[3] The three Gerundival Constructions have also been searched in the *MIDIA corpus* (Morfologia dell'Italiano in DIAcronia, Iacobini 2009), which consists of 7,652,526 words of written Italian from the beginning of the XIII century to 1947. The limited size of *MIDIA* and the sparseness of diachronic data (1223 tokens of *andare* + *gerund*, 203 tokens of *stare* + *gerund*, and only 152 tokens of *venire* + *gerund*) made it necessary to resort to the larger, although admittedly much noisier, *Google Ngram Corpus*. However, the lower timespan (1200-1947) and the more literary textual typology of this corpus in comparison with the *Google Ngram Corpus* has brought to different results: here the $\text{Continuous}_{\text{andare}}$ Construction is well attested, while the Progressive Construction is not so frequent.

[4] According to Bybee (2010: 67), in the case of argument structure constructions, productivity may be defined as the "likelihood that a construction will be extended to new items".

[5] Schematicity refers to the degree of differentiation within members of a category and it is index of the variation range within a class (Bybee 2010: 67). In the case of argument structure

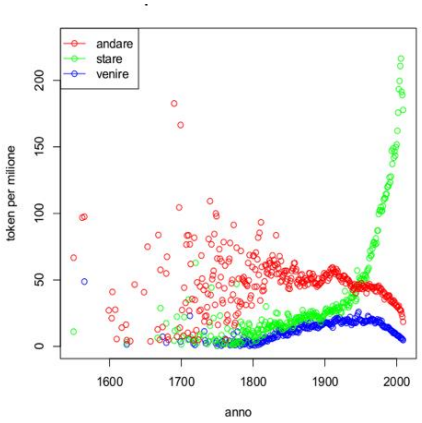

Figure 1. Constructions' Token Frequency.

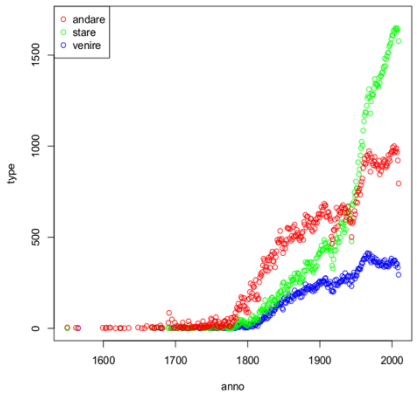

Figure 2. Constructions' Type Frequency.

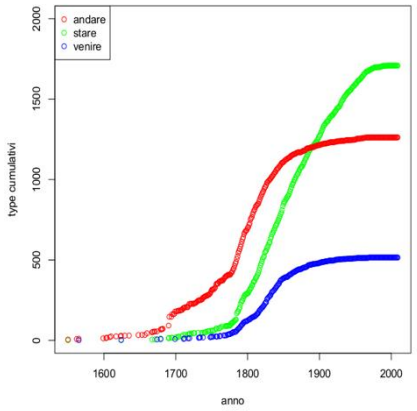

Figure 3. Constructions' Curve of Growth.

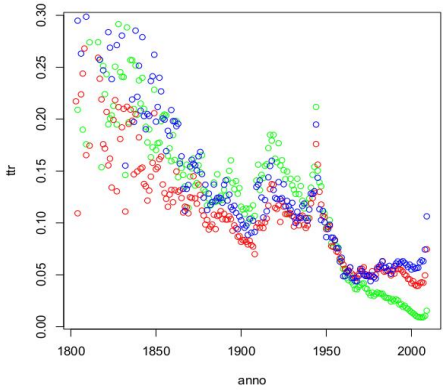

Figure 4. Constructions' Type/Token Ratio (1800-2000)[6].

4.1. *Andare* + gerund

This construction is undoubtedly the most represented one during the first centuries of Italian language. From the query performed on *MIDIA* (cf. note 2), whose results are not presented here, it is clear that this construction is well

construction, it concerns the classes of elements which may appear into the slot of a construction.

[6] In this case, we have depicted only the timespan 1800-2000 because the type/token ratio index is not informative before that time. Until the beginning of the XIX century, it is not uncommon to find type/token ratio values of 1: this result does not mirror the semantic variety of the constructions, since it depends on the very few occurrences of the periphrases in the *Google Ngram Corpus* before the XIX century.

represented since the XIII century (109 token per million and 76 types in 1200-1375, 143 token per million and 134 types in 1376-1532). This prominence is evident from *Google Ngram Corpus* too. As far as token frequency (cf. Fig. 1) is concerned, during the XIX century it appears twice more frequently than the others and its predominance lasts until the first decades of the XX century. After that, this construction undergoes a crisis, but it resists until the half of the century and it remains productive: even though its token frequency decreases, its type frequency (cf. Fig. 2) grows markedly around the half of the XX century. This phenomenon can be explained by the presence in the construction of many low frequency verbal types, which make it a periphrasis typical of a high linguistic register. Hence, it was the first construction to arise, then its crisis is marked by a reduction of type and token frequency, so that the phase of high productivity is to be located until the half of the XIX century. Despite being quite rare nowadays, the *andare* + *gerund* construction is still characterized by a good variety of medium-low frequency verbal types. Even the growth curve confirms the premature decrease in productivity, already identified by Bertinetto (1989-1990) and Squartini (1998).

4.2. *Venire* + gerund

In contrast with the expectations about a strong commonality between the two Continuous Constructions, corpus data have revealed an independent behavior of *venire* + *gerund*, so that a more fine-grained analysis seems essential to distinguish it from the Continuous$_{andare}$ Construction. Though as old as *andare* + *gerund* (as results from *MIDIA* confirm, cf. note 2), until the XVIII century available data are fairly scarce (cf. Fig. 1). It is however clear that the periphrasis has been growing in type and token frequency until the middle of the XX century and its productivity lasted longer than the one of *andare* + *gerund*. Interestingly, its token frequency is quite similar to the one of *stare* + *gerund* until the mid of the XIX century. Its crisis begins around the second post-war phase, reaching nowadays frequency values almost close to zero. Even though this construction is the least represented in the sample, it grew in the XX century too, but, since its values were very low, it has nowadays fallen into disuse (Giacalone Ramat 1995: 200). Type frequency is limited, but increases until the third quarter of the XX century. The type/token ratio (cf. Fig. 4) reaches its peak around the end of the XIX century. However, the growth curve (cf. Fig. 3) indicates a decrease of productivity at the end of the XIX century: this construction seems more varied than the others due to the presence of many low frequency types, almost idiomatic fossilized usages (such as many verba dicenda: *addurre* 'to adduce', *dimostrare* 'to demonstrate', *esporre* 'to expound').

4.3. *Stare* + gerund

The Progressive Construction is undoubtedly very infrequent until the XVIII century, but, during the following century, it undergoes a continuous growth, which become exponential within the second half of the XIX century (cf. Fig. 1). The lack of data for this construction in the first periods (in *MIDIA* too: 4 tokens per million and 5 types in 1200-1375, 6 tokens per million and 9 types in 1376-1532) is probably due to the fact that it was still in a primordial phase of its history. However, our sample suggests to anticipate its explosion to the beginning of the XIX century, instead of the second half of this century, as it has been proposed (Brianti 1992). Corpus data also confirm the great development of *stare* + *gerund* in the second post-war period, which lasts steadily until today. Looking at the type frequency (cf. Fig. 2), around 1930 the periphrasis becomes the most varied construction of the three, attracting more and more new verbs. The type/token ratio (cf. Fig. 4) reaches its peak in the 1900-1950 and the growth curve (cf. Fig. 3) maintains the highest rhythm of increase for a longer period, suggesting the strong expansion rate of this construction.

5. Distributional semantic analysis of the Gerundival Constructions

The change of the linguistic productivity of the Gerundival Periphrases has also been analyzed with distributional semantic methods. Distributional semantics is based on the so-called *Distributional Hypothesis*, according to which similarity of meaning correlates with similarity in distribution (Harris 1954, Lenci 2008). In distributional semantics, the meaning of words is estimated from the statistical analysis of their contexts, in a bottom-up fashion: requiring no sources of knowledge other than corpus-derived information about word distribution in contexts, thereby providing a usage-based model of meaning. Distributional semantics represents words as vectors built from their co-occurrences with linguistic contexts. The lexicon is thus modeled as a *semantic space* in which similarity between words is approximated in terms of the geometric distance between their vectors. Distributional semantic spaces are usually built with a four-step method (Turney and Pantel 2010): for each target word, contexts are collected and counted and a co-occurrence matrix is generated; raw frequencies are then usually transformed into significance scores that are more suitable to reflect the importance of the contexts; the resulting matrix tends to be very large and sparse, requiring techniques to limit the number of dimensions. Finally, a similarity score is computed between the vector rows, using various similarity measures.

We have used distributional methods to investigate the changes undergone by the semantic space of the verbs occurring in the Gerundival Constructions throughout the period represented by our corpus data. In fact, the productivity of a construction does not only depend on the sheer number of different items occurring with it, as measured by type frequency, but also on the semantic diversity of such items (Barðdal 2008; Bybee 2010; Suttle and Goldberg 2011). Therefore, a highly productive construction instantiates a high number of very dissimilar items. To this purpose, we have represented each verb type appearing in the gerundival constructions with a distributional vector, built by extracting their co-occurrences with the top 30,000 content words from *La Repubblica* and *Paisà* Italian corpora. The co-occurrence counts were collected with a context window of ± 5 content words from each target word. The obtained matrix was then weighted by Positive Local Mutual Information (PLMI) and reduced to 300 latent dimensions via Singular Value Decomposition (SVD).

Since we use distributional data for a diachronic analysis, the optimal way to proceed would have been to build a distinct vector space for each temporal window using only distributional data coming from texts belonging to that period. However, the scarce documentation of the first centuries in the *Google Ngram Corpus* is not sufficient to build reliable distributional spaces, which notoriously suffer from data sparseness. Therefore, following Perek (2016), we adopted the simplifying solution of using a single distributional space trained on data coming from corpora of contemporary Italian. Of course this procedure is legitimate only by assuming that the meaning of the analyzed verbs is not likely to have changed considerably within the time frame considered in this survey.

The organization of the semantic space of the verbs appearing with a given construction was analyzed through *Multidimensional Scaling (MDS)*, a statistical technique that allows to visualize the associations that are contained in the data, ignoring the less informative components (Jenset 2014: 9-10). MDS assigns to distributional vectors their coordinates in a two-dimensional space according to their similarity relations. We have divided Google data into eleven groups, favoring a more fine-grained division of twenty-year frames for the last periods: C (1550-1691), D (1692-1840), E (1841-1861), F (1862-1884), G (1885-1905), H (1906-1926), I (1927-1947), L (1948-1968), M (1969-1989), N (1990-2000), O (2001-2009)[7]. For each period, the verbs occurring with a given con-

[7] This grouping has been based on the division of the *MIDLA* corpus (A: 1200-1375, B: 1375-1532, C: 1533-1691, D: 1692-1840, E: 1841-1942) in order to make data from different sources comparable.

struction in that period have been plotted using the coordinated produced by MDS. Therefore, it is possible to follow the pattern of semantic development of a construction by analyzing the new verbs it attracts during time. At a local level, plot areas that are more densely populated by new verbs can be considered the most productive ones.

Despite the dimensions of MDS spaces do not come with an explicit label, the axes in the verb space seem to capture relevant semantic dimensions. The y-axis represents a variation in concreteness: more physical actions (induced movement, creation, action on a patient: *gettare* 'to throw', *distruggere* 'to destroy', *costruire* 'to build') appear in the upper part of the graphic, whereas abstract actions (semantic fields of speech, thought, persuasion: *domandare* 'to ask', *immaginare* 'to imagine', *sollecitare* 'to urge') are easily identifiable at its bottom. The x-axis is instead harder to interpret, but telicity, transitivity and agentivity increase towards the right side of the plot, with verbs expressing actions aimed towards a goal, verbs that require a patient object to carry out the action on, verbs of transformation, etc. (e.g., *fabbricare* 'fabricate', *pubblicare* 'to publish'). It is also possible to notice the presence of verbs that identify processes of creation, aggregation, organization, fulfillment (such as *completare* 'to complete', *associare* 'to associate', *realizzare* 'to achieve'). On the left side, verbs expressing violence, disintegration, dispersion, movement and alteration are to be found (for instance, *scomparire* 'to disappear', *imprecare* 'to swear', *cascare* 'to fall', *avvelenare* 'to poison').

In the following subsections we present and discuss the diachronic semantic plots (Figures 5-32) of the verbs appearing with the three Gerundival Constructions. Verbs that had already appeared with a construction in any previous phase are marked as red points. On the other hand, blue points refer to verbs that appear for the first time in a construction in that specific temporal frame.

5.1. *Andare* + gerund

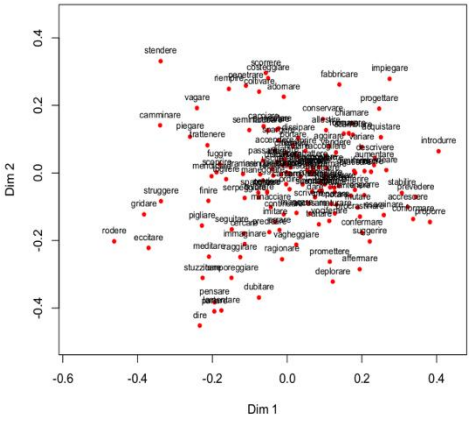

Figure 5. Semantic space of andare + *gerund (1550-1691).*

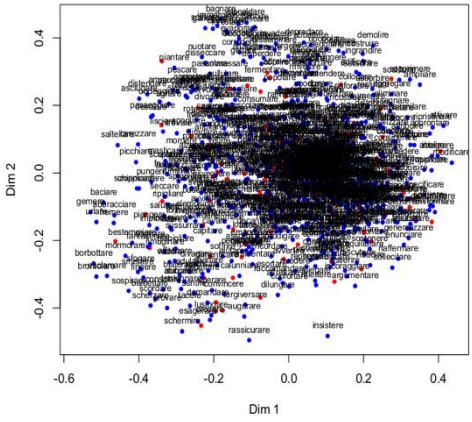

Figure 6. Semantic space of andare + *gerund (1692-1840).*

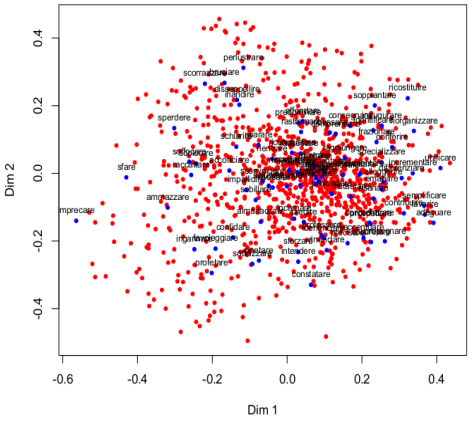

Figure 7. Semantic space of andare + *gerund (1841-1861).*

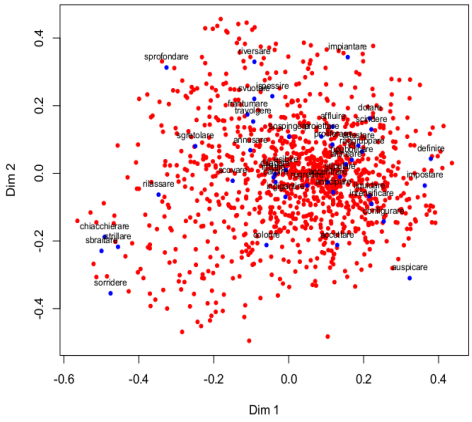

Figure 8. Semantic space of andare + *gerund (1862-1884).*

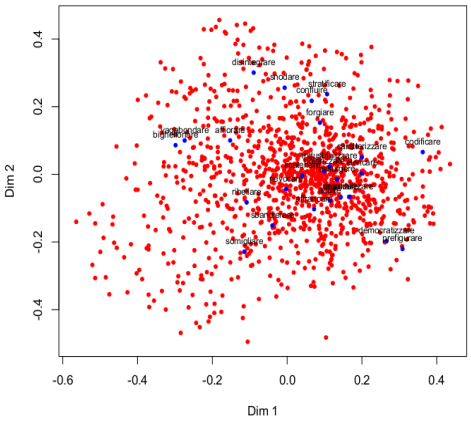

Figure 9. Semantic space of andare + *gerund (1885-1905).*

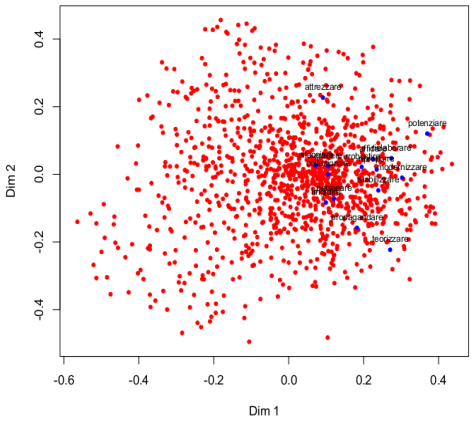

Figure 10. Semantic space of andare + *gerund (1906-1926).*

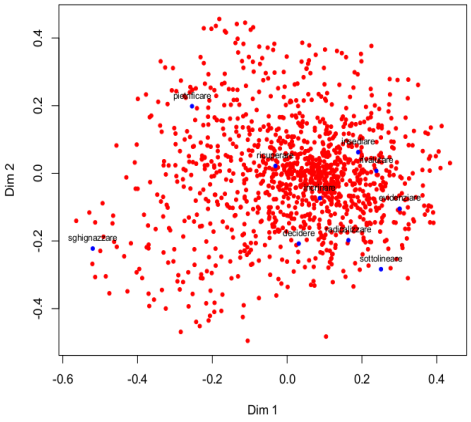

Figure 11. Semantic space of andare + *gerund (1927-1947).*

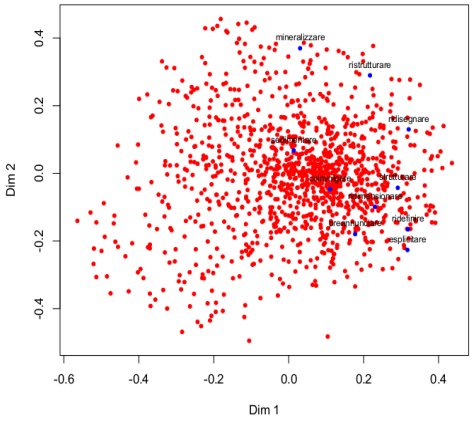

Figure 12. Semantic space of andare + *gerund (1948-1968).*

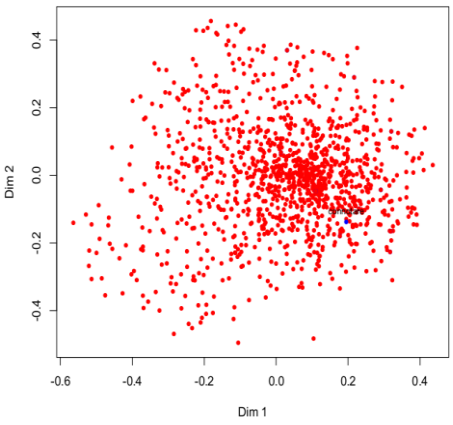

Figure 13. Semantic Space of andare + *gerund (1969-1989).*

First of all, we can observe a general bias of this construction towards high telicity and transitivity, identifiable at the right center of the plot (cf. verbs of physical action on an object, such as *accendere* 'to switch on', *ampliare* 'to expand', *rompere* 'to break'). The most prominent growth is located in the main area of the plot, which is the first to be filled up. In the period 1692-1840, extensions spread radially from here, shaping a cloud of novel uses. This gradually diminishes its density moving from the center to the periphery and then goes beyond the semantic borders marked by red points, which stand for those verbs that have already appeared in a previous period. In the second half of the XIX century productivity begins to decrease and a few novel items shape small clusters composed by two or three verbs. The expansion towards intransitive verbs is now concluded, whereas new telic verbs, which identify transformations performed on an object patient, are still to be found in this periphrasis. In the XX century, some new isolated telic verbs appear, probably as item-based analogies with previous existing semantic areas. They identify mostly dialectical or intellectual activities aimed at a goal (*sottolineare* 'to highlight', *ridimensionare* 'to scale down', *preannunciare* 'to predict', *connotare* 'to connote').

Regarding the semantic types of the involved verbs, the main function of this construction seems to have been the expression of the following concepts: fulfillment of a goal-oriented process (cf. telic verbs), generic transitivity often associated with concrete actions, modality of intransitive actions (cf. verbs related to emotions). When productivity decreases, single isolated usages arise mainly in the area of high telicity, which remains productive, and often within semantic fields similar to those of the Continuous$_{venire}$ Periphrasis. It can hence be concluded that once this construction was somehow complementary to *venire* + *gerund*, as their different development proves (cf. statistical indices and semantic analyses). Conversely, nowadays they tend to share most contexts, while the construction *stare* + *gerund* has progressively occupied the original space of the *andare* + *gerund* construction, thanks to its greater extendibility and semantic neutrality. The Progressive Construction has become the unmarked mean to express imperfectivity because it does not contain information on the modality of the action and has no deictic orientation. Before that, the Continuous$_{andare}$

Construction was the only argument structure construction available to express imperfectivity in addiction to incrementality and modality, with the deictically oriented variant Continuous$_{venire}$ Construction, less frequent.

5.2. *Venire* + gerund

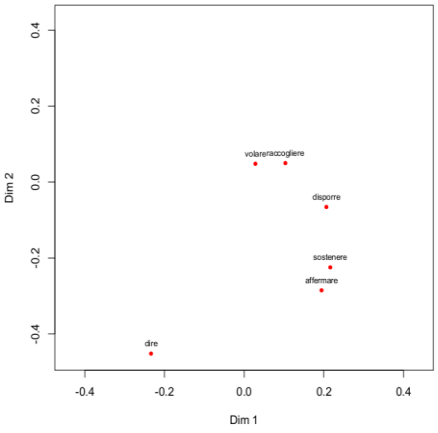

Figure 14. Semantic space of venire + *gerund (1550-1691).*

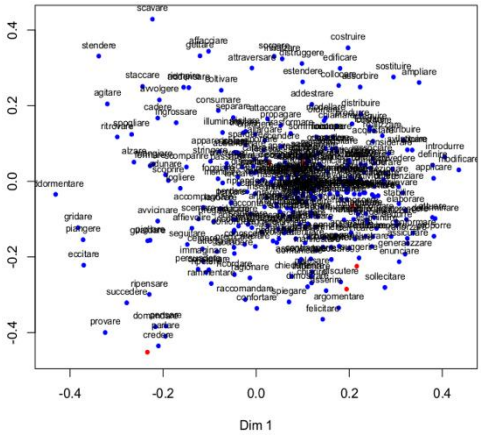

Figure 15. Semantic space of venire + *gerund (1692-1840).*

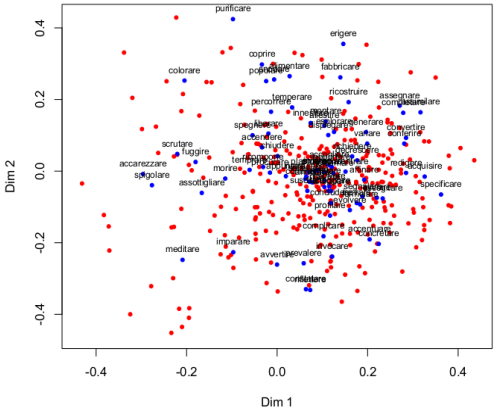

Figure 16. Semantic space of venire + *gerund (1841-1861).*

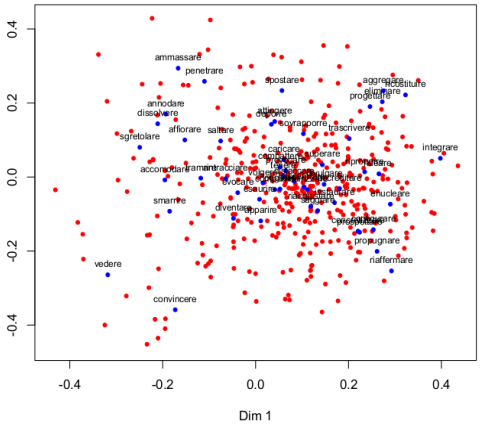

Figure 17. Semantic space of venire + *gerund (1862-1884).*

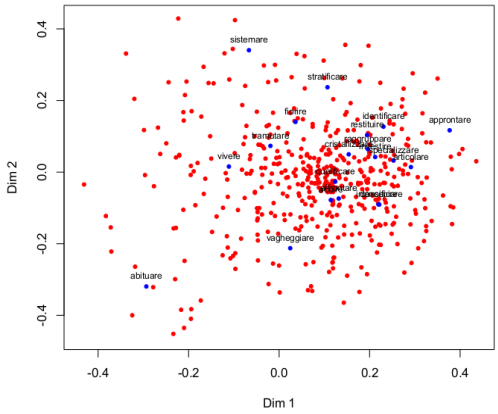

Figure 18. Semantic space of venire + *gerund (1885-1905).*

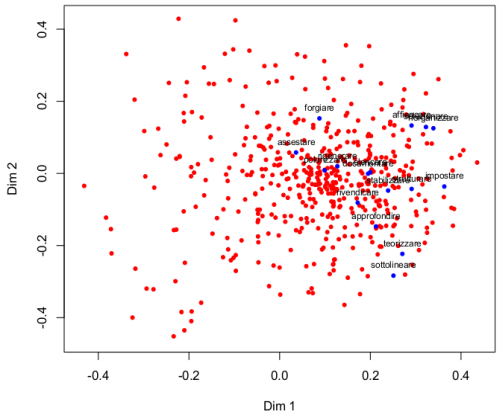

Figure 19. Semantic space of venire + *gerund (1906-1926).*

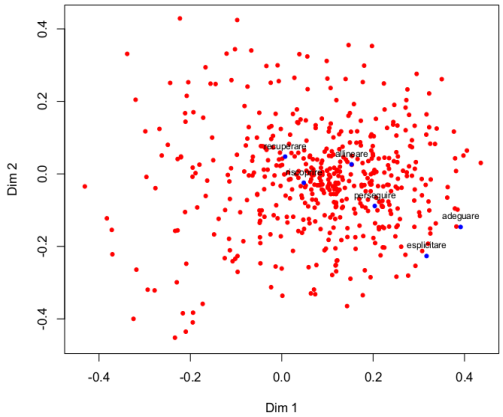

Figure 20. Semantic space of venire + *gerund (1927-1947).*

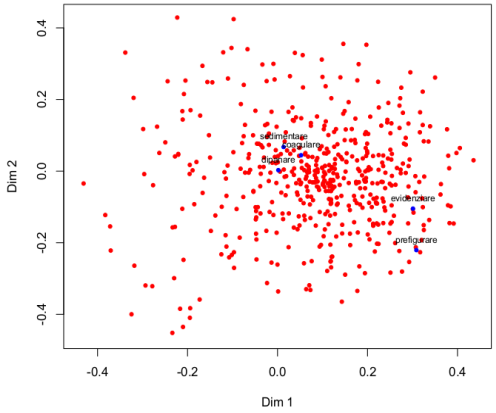

Figure 21. Semantic space of venire + *gerund (1948-1968).*

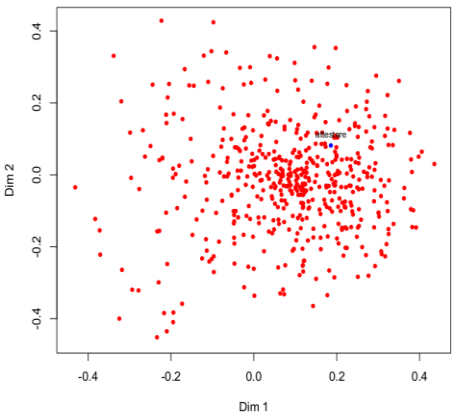

Figure 22. Semantic space of venire + *gerund (1969-1989).*

The semantic space of this construction starts to get crowded in the XVIII century. In this period, a dense group arises at the right center of the plot. Here is where the verbs that identify formation, ordering and fulfillment processes are located (e.g., *collocare* 'to situate', *acquistare* 'to acquire', *modificare* 'to modify'). Nevertheless, there are also some extensions (intransitive, such as *scomparire* 'to disappear', *sorgere* 'to arise') on the left side. The second half of the XIX century is the last really productive period for this construction. The main area attracts telic and transitive verbs, mainly actions of control on an object, verbs of discussion and events related to an idea of order or formation (for instance, *plasmare* 'to shape', *concludere* 'to conclude', *specificare* 'to specify', *conferire* 'to confer', *chiudere* 'to close', *aggregare* 'to aggregate', *strutturare* 'to organize', *proporre* 'to propose'). This is the only area that remains productive over the following century, even if the new extensions become isolated and analogically modeled after single examples. Due to the marked preference for telic verbs, especially those of completion, ordering or formation processes, it appears legitimate to assert that the main function of this construction is that of conveying a process from the point of view of its fulfillment.

5.3. *Stare* + gerund

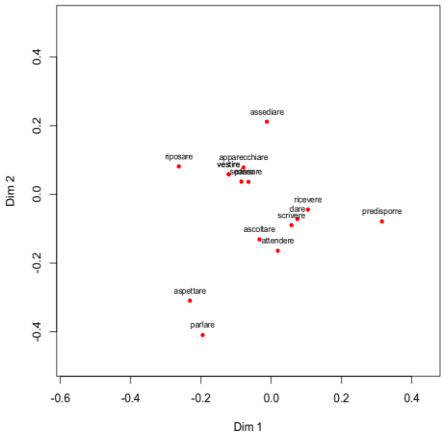

Figure 23. Semantic space of stare + *gerund (1550-1691).*

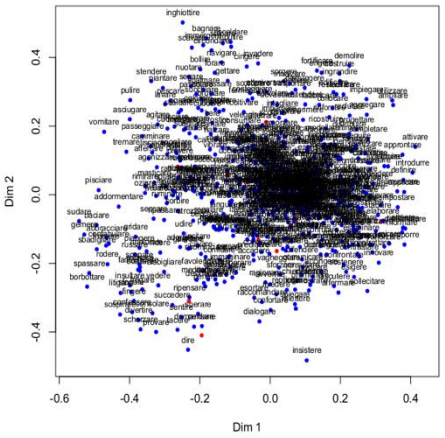

Figure 24. Semantic space of stare + *gerund (1692-1840).*

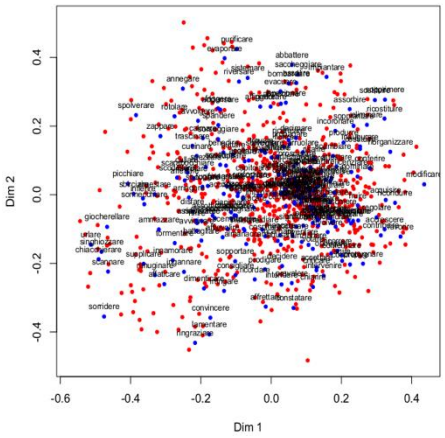

Figure 25. Semantic space of stare + *gerund (1841-1861).*

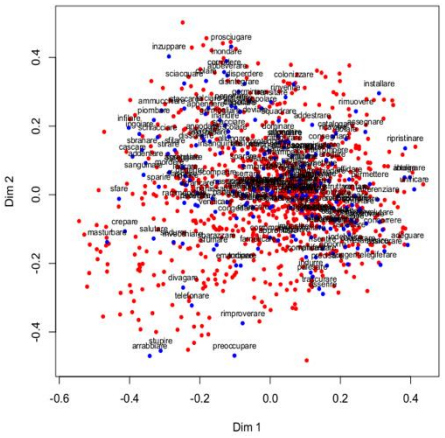

Figure 26. Semantic space of stare + *gerund (1862-1884).*

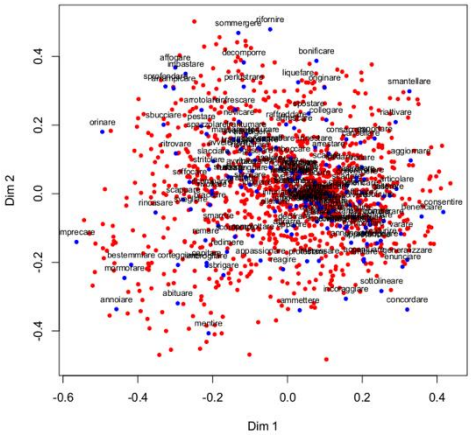

Figure 27. Semantic space of stare + *gerund (1885-1905).*

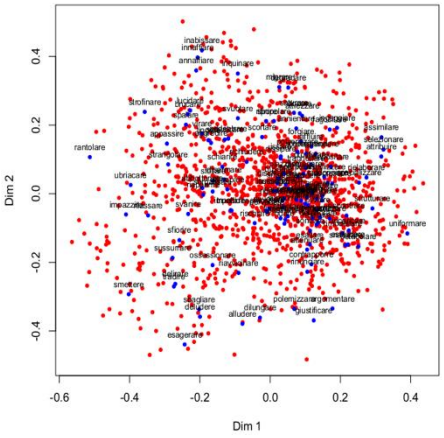

Figure 28. Semantic space of stare + *gerund (1906-1926).*

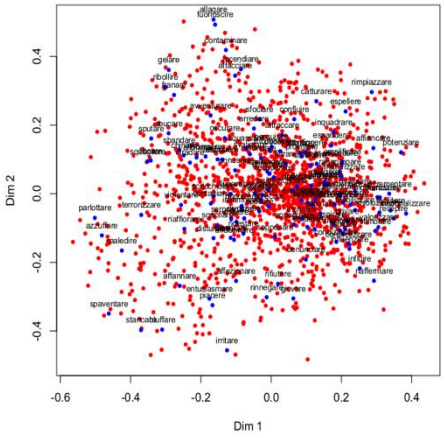

Figure 29. Semantic space of stare + *gerund (1927-1947).*

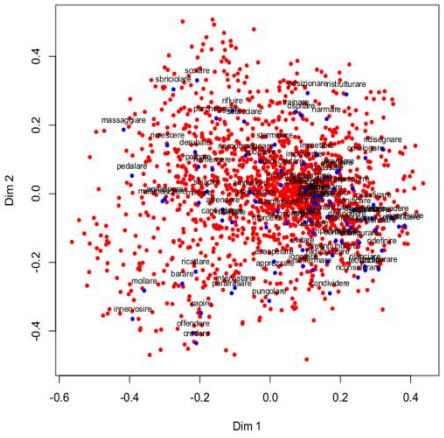

Figure 30. Semantic space of stare + *gerund (1948-1968).*

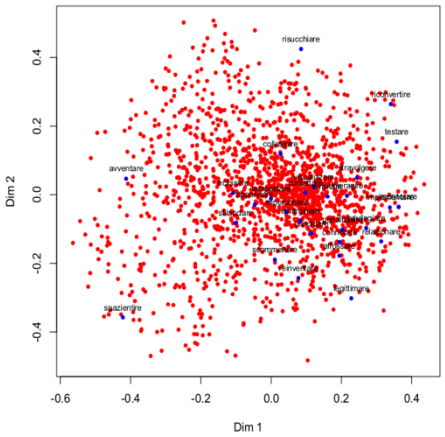

Figure 31. Semantic space of stare + *gerund (1969-1989).*

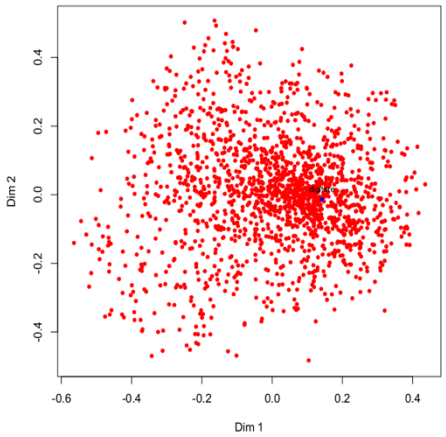

Figure 32. Semantic space of stare + *gerund (1990-2000).*

A main verb group arises in the XVIII-XIX centuries in the center of the plot. This set is organized in many small sub-clusters that refer both to abstract and to concrete actions (*sentire* 'to feel', *ricevere* 'to receive', *passare* 'to pass', *compiere* 'to accomplish', *scrivere* 'to write', *comprare* 'to buy'). In the second half of the XIX century, the strength of attraction of this group increases, whereas the outer portions of the plot stop being productive, so that the construction becomes almost semantically neutral and more compact. Most of the verbs indicate deep modifications of an object, sometimes in a violent manner (*sconvolgere* 'to upset', *deviare* 'to deviate', *sparare* 'to shoot', *disintegrare* 'to disintegrate'). Instead, peripheral analogies lead to the rise of new items related to the semantic field of emotion or elocution (*ridere* 'to laugh', *fremere* 'to quiver', *borbottare* 'to grumble'), which expand with less energy. The period 1900-1920 is marked by great productivity: new extensions cover the whole semantic space, growing radially from the center, which does not represent the only attractor any more. In the following period, the amount of new elements reduces, but they are still located in the area of highest transitivity and telicity. The construction seems to be mostly associated to durative verbs in general, without being limited to any specific semantic field (some examples: *convergere* 'to converge', *peggiorare* 'to worsen', *proliferare* 'to proliferate', *risparmiare* 'to spare', *relazionare* 'to relate'). This great semantic diversity is another clue of the strong productivity of such a construction. Preference for telicity tends to increase in time, even if the construction is not particularly developed in this direction.

The construction keeps growing considerably, until the second half of the XX century, towards many different directions. Preferred verbs refer to con-

crete actions and have a general meaning, as it is typical of a very productive pattern. This great expansion occurs during the same time frame in which the growth of the construction *andare* + *gerund* starts decreasing. Moreover, the semantic spaces covered by the instances of both constructions are quite similar. These two phenomena seem to indicate a replacement of the Continuous$_{andare}$ Construction by the Progressive Construction.

5.4. Measuring the semantic density of the constructions

The cosine of the angle between two vectors is a common measure of their similarity (Lenci 2008). Given a set of words, their average cosine similarity can be used to measure the *density* of the semantic space covered by this set: a high average cosine in fact means that the words are grouped close together, while a low average cosine means that they are very spread apart. The low semantic density of a category entails that its members are semantically diverse, and this in turn is an important factor of productivity. Highly productive constructions can be applied to items that are very spread in the semantic space. Therefore the set of items occurring with a highly productive construction will have a lower semantic density than the items of a less productive construction, which can expand only to new items that are analogically very close to old items (Suttle and Goldberg 2011; Perek 2016).

For each Gerundival Construction and for each time period, we computed the average cosine of their verbs, as an estimate of the semantic density of the construction for that period. We predict that the productivity change of the constructions corresponds to a change in their semantic density, with the constructions becoming less productive increasing the average cosine of their verbs.

Regarding the two Continuous Constructions, their average cosine becomes higher and higher (Fig. 33), showing how the verbs, which appear in these periphrases, tend to become always more cohesive one another. An increasing cosine should indicate the development of highly dense portions of space, which may easily attract new usages. However, their productivity's decrease, which this study already introduced, proves that they tend not to expand. Indeed, the presence of new verbs is due to item-based analogical formations, which increase the average cosine. Semantic space reduces in extension and contains few clusters, which become denser and denser.

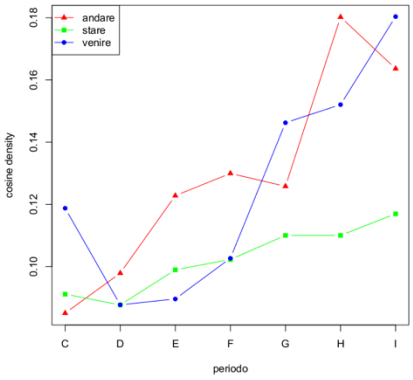

Figure 33. Average cosine for the verbs in the Constructions.

On the other hand, the verbs that appear in the Progressive Construction show a lower average cosine. This fact reflects the high productivity of the construction, especially starting from the period E (1841-1861). Verbs spread out all over the semantic space, as it is appropriate for a highly productive construction that produces a high amount of diverse usages.

5.5. The distributional analysis of construction productivity: some general remarks

The diachronic analysis of the distributional semantic space has shown that the verbal group containing the majority of elements at the beginning is always the one which grows faster and attracts the highest amount of new items. This behavior confirms the essential role of type frequency for productivity. As Perek (2016: 21) underlines: "within a given semantic class, new items are more likely to appear if many items are already present". However, productivity depends not only on the previous experience of speakers with the construction, but also on the coverage of the novel element by the already existing semantic space. As this study shows, new usages tend to be located within the limits of the semantic space that has already been identified in previous periods. The more populated a semantic area is, the more likely new usages are.

Furthermore, the extensibility of argument structure constructions largely depends on item-based analogy. An element with high token frequency is more likely to become an analogical model and to attract new similar items in the syntactic pattern in which it recurs. So, token frequency influences the possibility of analogies and, consequently, the construction's productivity. Indeed, type frequency is an essential index of productivity for highly variable and schematic constructions (cf. *stare* + *gerund*), but it is not a sufficiently adequate factor in the case of semantic coherent constructions, namely constructions attested only with semantically similar types. For the latter, semantic similarity represents a better criterion of productivity.

This confirms the view of productivity proposed by Barðdal. Syntactic productivity is a function of construction's type frequency, semantic coherence

and the inverse correlation between the two (2008: 34). Because of these relations, productivity is gradient and constructions are located on different positions on a schematicity scale (id.: 36). An argument structure construction, which has been experienced with few different verbs, will be productive only if the acceptable verbs are similar one another, because novel usages are admitted only within the restricted semantic domain defined by these items. On the other hand, a construction that admits very different verbs needs a high type frequency to be productive, because semantic variability has to be attested by a sufficient amount of elements in order to attract new items.

The two Continuous Constructions seem to be a case of low type frequency and high semantic coherence; this is why their expansion occurs according to analogical processes. On the contrary, the Progressive Construction is characterized by high type frequency and limited semantic coherence, thus its expansion is mostly driven by the slot schematicity. Its high type frequency, the low average cosine (i.e., semantic density) and the way in which it attracts new verbs make this periphrasis a prototypical example of schema-based productivity. It is also remarkable that, during its first period, the Continuous$_{andare}$ Construction was characterized by a great schema-based productivity, but this progressively changed into an analogy-driven process. In contrast, *venire* + *gerund* has always been based on semantically well-defined clusters of verbs, thus representing an example of low frequency and high semantic coherence construction, even though with a lower level of productivity in comparison to the other Continuous Construction.

6. Conclusions

Thanks to statistical and distributional analyses, we have provided new quantitative data about the diachronic development of the three Italian Gerundival Constructions. The reliability of the results is guaranteed by the large amount of data automatically extracted from the *Google Ngram Corpus*. To the best of our knowledge, this is the first time this corpus has been used to explore the evolution of construction productivity, thereby confirming its potentialities for diachronic analysis.

First of all, apart from the single results concerning each periphrasis, the aim of this study was to combine, in a diachronic perspective, two different approaches, Construction Grammar and Distributional Semantics. In the previous chapters, we have highlighted the adequacy of Construction Grammar to interpret the dynamics of linguistic objects and explore the determinants of productivity. Moreover, we have confirmed the suitability of Distributional Semantics to analyze the evolution of constructional meaning in diachrony. By

considering the meaning as gradient and continuous, it has been possible to introduce a dynamic perspective to language change, which can consequently be seen as a sequence of objective distributional changes. The results of our analysis support the importance of adopting these two combined approaches.

The evolution of the Gerundival Periphrases with respect to the verbal items they occur with shows that construction productivity depends on the previous experience of the speakers with the pattern, consistently with the usage-based paradigm shared by Construction Grammar and Distributional Semantics: the more elements have already co-occurred with a construction, the more likely it is that new items will appear in it. Moreover, constructions also expand through item-based analogies, which stem from high similarity of new items with specific old instances. These have already been produced by the speakers, and, in time, they become analogical models by attracting elements which are semantically related to them. Moreover, constructions exist at different levels of schematicity. On one hand, there are extremely variable and schematic constructions, marked by high type frequency and low semantic coherence between types, which expand thanks to the schematicity of their slots. A good example of this is the *andare* + *gerund* construction in its old stages, and the *stare* + *gerund* construction in its recent evolution. On the other hand, there are semantically coherent constructions, generally characterized by low type frequency, but high semantic similarity between single instances, which expand through local analogical patterns. The current stages of the Continuous Constructions are a clear example of this type of process.

As a final remark, our study confirms the rich potentialities coming from computational analyses of large corpus data in synergy with usage-based models of language, to gain new insights on the patterns of lexical and grammatical change.

References

Barðdal, J. (2008), *Productivity. Evidence from Case and Argument Structure in Icelandic*, Amsterdam/Philadelphia, John Benjamins Publishing Company.

Bertinetto, P.M. (1989-1990), *Le perifrasi verbali italiane: saggio di analisi descrittiva e contrastiva*, in *Quaderni patavini di linguistica* 8-9, pp. 27-64.

Bertinetto, P.M. (1997), *Il dominio tempo-aspettuale. Demarcazioni, intersezioni, contrasti*, Torino, Rosenberg & Sellier.

Brianti, G. (1992), *Périphrases aspectuelles de l'italien. Le cas de* andare*,* venire *et* stare + *gérondif*, Berne, Peter Lang.

Bybee, J. (2010), *Language, usage and Cognition*, New York, Cambridge University Press.

Giacalone Ramat, A. (1995), *Sulla grammaticalizzazione dei verbi di movimento*, in *Archivio Glottologico Italiano* 80, pp. 168-203.

Goldberg, A.E. (1995), *Constructions. A Construction Grammar approach to Argument Structure*, Chicago and London, The University of Chicago Press.

Hoffman, T., Trousdale, G. (2013), *The Oxford Handbook of Construction Grammar*, New York, Oxford University Press.

Jenset, G.B. (2014), *Mapping meaning with distributional methods: a diachronic corpus-based study of existential there*, in *Journal of Historical Linguistics* 3(2), pp. 272-306.

Lenci, A. (2008), *Distributional semantics in linguistic and cognitive research*, in *Italian Journal of Linguistics* 20(1), pp. 1-31.

Perek, F. (2016), *Using distributional semantics to study syntactic productivity in diachrony: A case study*, in *Linguistics* 54(1), pp. 149-188.

Squartini, M. (1990), *Contributo per la caratterizzazione aspettuale delle perifrasi italiane* andare + *gerundio,* stare + *gerundio,* venire + *gerundio. Uno studio diacronico*, in *Studi e saggi linguistici* 30, pp. 117-212.

Squartini, M. (1998), *Verbal Periphrases in Romance*, Berlin, Mouton de Gruyter.

Suttle, L., Goldberg, A.E. (2011), *The partial productivity of constructions as induction*, in *Linguistics* 49, pp. 1237–1269.

Turney, P.D., Pantel P. (2010), *From Frequency to Meaning: Vector Space Models of Semantics*, in *Journal of Artificial Intelligence Research* 37, pp. 141-188.

Finito di stampare nel mese di marzo 2017
presso Tipografia Monteserra S.n.c. - Vicopisano
per conto di Pisa University Press