The diachrony of Italian word formation through the Google Books corpus: NN compounds and their rival syntagmatic NPN counterparts

M. Silvia Micheli[1] Jan Radimský[2] Tomáš Mrkvička[2] Pavel Štichauer[3]

[1]University of Milan

[2]University of South Bohemia, České Budějovice

[3]Charles University, Prague

Biennial of Czech Linguistics, 17-20 September 2024

Outline

1. Introduction: competition in morphology
2. NN and NPN in Italian: key properties
3. Research questions
4. Methodology
5. Detailed results: case studies
6. General trends: clustering
7. Conclusions

Introduction: competition in morphology

- **Disclaimer: a first version of this presentation was aired at the IMM21 (Vienna, 28-30 August 2024); this version brings a novelty... (clustering performed by Tomáš Mrkvička)**
- **Competition** between morphological strategies has recently been the subject of renewed interest (see Rainer et al., 2019 and Masini, 2019b).
- Nevertheless, while the prevailing focus has been on synchronic data, the extent to which **diachronic perspectives** can shed light on the emergence and competition of morphological constructions remains unclear.
- In this study, we explore the diachronic development of two constructions that are currently regarded as competitors (Masini, 2019a):
  - **Noun-Preposition-Noun phrasal nouns** (e.g., *sala da tè* 'tearoom')
  - **Subordinate Noun-Noun compounds** (e.g., *sala stampa* 'press room')

NN and NPN in Italian: key properties

- Italian NPN phrasal nouns are lexical constructions "that are formally akin to phrases, but lexical in nature" (Masini & Scalise, 2012).
  - They have a clear **naming function**
  - **Paradigmatic variation is blocked** (the nouns cannot be substituted by near-synonyms)
  - The internal members of the NPN phrasal noun **cannot be interrupted by adjectives or adverbs**
- They are made up of two nouns separated by a **preposition**
  - According to the GRADIT dictionary, *DI*, *A* and *DA* are the most commonly used prepositions, whereas *IN*, *CON*, *PER* and *SU* are attested but more limited (Masini, 2009, p. 259).

NN and NPN in Italian: key properties

- The second noun (N2) can be both **bare** and preceded by a **determiner** (DET)
  - *casa di cura* 'nursing home' N-PREP-N
  - *casa dello studente* 'student house' N-PREP+DET-N
- As reported by Masini (2009, p. 261), in the GRADIT dictionary, phrasal nouns with bare N2 are more frequent than those with a determiner

Table 1: N+PREP+N *versus* N+PREP+DET+N phrasal nouns in GRADIT (Masini, 2009, p. 261)

Bare PREP	**Figures**	**PREP+DET**	**Figures**
di 'of'	10252	*di*+DET	3843
a 'at'	1986	*a*+DET	397
da 'from'	728	*da*+DET	189
in 'in'	311	*in*+DET	28
per 'for'	204	*per*+DET	41
con 'with'	33	*con*+DET	9
su 'on'	24	*su*+DET	30

NN and NPN in Italian: key properties

- **NPN phrasal nouns in Italian** (and in other Romance Languages such as French, see Goethem and Amiot, 2019) have been extensively investigated
- Traditionally assimilated to compounds (see, e.g. Benveniste, 1966 and Tollemache, 1945, respectively)
- Masini (2009, 2019a) explores Italian NPN phrasal nouns and compounds within Construction Morphology, analyzing their continuum and synchronic competition: *"competition operates* ***in different directions*** *(words may block possible MWEs, but also MWEs may block the formation of complex words) and possibly at different levels of abstraction, since it may involve specific lexical items but also patterns of formation"*

Subordinate NN compounds: key properties

- **Subordinative NN compounds (Radimský, 2015)**
  - Productive pattern that forms complex naming units
  - Involves 2 bare common nouns (no determiner)
  - Implicit relationship between nouns (no preposition)
  - Order of constituents: mostly endocentric, left-headed
    - *trattamento rifiuti* (treatment-wastePL) – waste treatment
    - *trattamento rifiuti È UN trattamento* (waste treatment IS A /kind of/ treatment)
- **Two types:**
  - **Verbal nexus compounds, VNX**: deverbal head + Non-head element interpreted as its argument (*controllo passaporti* 'passport control')
  - **Grounding compounds, GRD**: other kind of subordinate relationship (*pausa pranzo* 'lunch break')

Research questions

- In this study we examine a sample of subordinate NN compounds and their **potential or existing NPN equivalents**, based on data extracted from the Google Books corpus
  - we provide an overall description of the competition between NN and NPN (both simple PREP and PREP+DET)
  - we focus on the competition between VNX NN and NPN with DI / DI+DET preposition in diachrony
- Our aim is to answer the following **research questions**:
  1. Which NN types are most in competition with NPN constructions?
  2. Are there NNs that do not have an NPN counterpart?
  3. Are there **'niches'** where the NN pattern gradually dominates over the NPN pattern?
  4. How does the competition between VNX NNs and NPNs with preposition *DI* evolve in diachrony?

Data gathering

- The study is based on extensive diachronic data drawn from the Google Books corpus (size: 120,410,089,963 tokens) available in the form of raw frequency lists - Data for the extraction of NNs and NPNs come from pre-treated bigrams and trigrams
- We extracted a sample of **4,230 SUB NNs** - **2,458 GRD (grounding)** and **1,772 VNX (verbal-nexus) NNs**
  - **Manual filtering**: based on previous research (Radimský, 2015), N1 and N2 families, N2 modifiers listed by the Zingarelli dictionary
  - **Manual verification of NNs** in Google Books in order to achieve a higher accuracy (many false positives have been eliminated)
  - **Manual verification of NPN competitors** (if present) in Google Books
    - Prepositions: *DI*, *A*, *DA* with(out) an article
    - Focus on the preposition *DI*/*DI*+Article – by far the most frequent in data

Summary statistics

- For each NN/NPN, dated numbers of occurrences in Google books are available from 1900 to the present with a year-by-year precision
- The number of tokens was set to a threshold of 40
- This allows us to analyse:
  - The overall presence or absence of NPN competitors in the sample, as in the table in Figure 1
  - The rate of NNs compared to NP(Art)Ns in diachrony

GRD	DiBare_no	DiBare_yes	Total:	DiBare_no	DiBare_yes	Total:
DiDet_no	1134	439	1573	46,1%	17,9%	64,0%
DiDet_yes	343	542	885	14,0%	22,1%	36,0%
Total:	**1477**	**981**	**2458**	60,1%	39,9%	100,0%

VNX	DiBare_no	DiBare_yes	Total:	DiBare_no	DiBare_yes	Total:
DiDet_no	~~333~~ 144	269	602	~~18,8%~~ 8,1%	15,2%	34,0%
DiDet_yes	166	1004	1170	9,4%	56,7%	66,0%
Total:	**499**	**1273**	**1772**	28,2%	71,8%	100,0%

Figure 1: Overall presence of NPN competitors in the sample (number of types)

NPN competitors in the sample

- **NN types with no NPN competitors (P = *di* or *di* + article)**
  - GRD: 46,1% (1134 types)
  - VNX: 18,8% (333 types), but at a closer inspection only 8,1% (144 types)
  - Half of them would require another preposition (*abbonamento a Internet* – 'Internet subscription') or are quasi-coordinative result nominals with N1 *aiuto* (*aiuto cameriera* – 'waitress assistant')
    - N1/N2 families with highest type frequency of NNs without competitors (no. of types):
    - N1: *rischio* (8) - 'risk', *assistenza* (7) - 'assistance', *deposito* (5) - 'deposit', *trattamento* (4) - 'treatment'
    - N2: *auto* (7) - 'car', *video* (5) - 'video', *hardware* (4) - 'hardware', *bagagli* (3) - 'luggage'
- **Results surprisingly consistent with data from present-day corpora**: cf. Baroni et al. (2009) for Italian, Radimský (2020) for French

Core of competition: Verbal nexus NNs x NP(Art)Ns

- **Focus on prototypical VNX NNs – 1360 NN types**
  - Such as *raccolta dati* - *raccolta di dati* - *raccolta dei dati* ('data collection')
  - Deverbal head + direct object; head ≠ *aiuto*, *aiuti* ('assistant')
  - Only 8,1% (144) of types without competitors with the preposition *di*
  - Such an NP(Art)N competitor is always available in theory
  - High rate of both NPN / NPArtN competitors – 56,7% (1004) types
  - No need to make a distinction between NPNs and NPArtNs
- **Methodology: analysis of the rate of NNs over NPNs / NPArtNs in diachrony**
  - E.g.: raccolta dati - raccolta di dati - raccolta dei dati ('data collection') – (near) synonyms, "overabundance"
  - Percentage of NN (*raccolta dati*) of these three variants?

Example: Relative token frequencies (GN viewer) vs. Rate of NNs

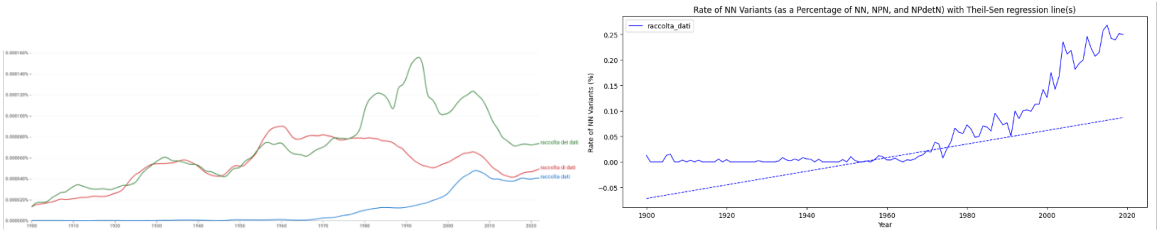

Figure 2: GN viewer for 'data collection' *versus* Rate of NNs over time

- Rate of NNs
  - % (range: 0,00 – 1,00)
  - Makes it possible to compare units with different frequencies and with changing fq. over time
  - The rate of the NN variant increases since 1970s to 25%

Results: individual NNs/NPNs

- Rate of NNs - a surprising variety of curves, including extreme cases:
  - *rassegna stampa* ('press review') → increase from 0 to almost 100
  - *allevamento cavalli* ('horse breeding') → decrease from 100% to almost 0%
  - *trasporto merci* ('freight transports') → increase/decrease: NN popular in 1940's with a rate of 67%

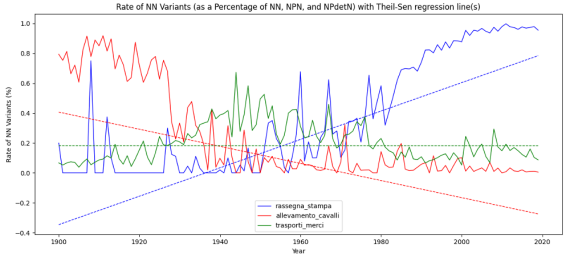

Figure 3: Rate of selected NNs over time

Results: do many NNs win in diachrony?

- **Range of proportions of NNs divided in 6 periods of 20 years** (see the following Table 2)
- **"Overabundance" scale** adapted from Thornton (2019) (majority + marked merged → majority)
- **Observations and conclusions:**
  - The competition progressively takes place: from NPN dominant to NPN majority
  - Very few diachronic NN winners – provided that almost all NNs emerged only after 1900's
  - NNs and NPNs mostly coexist in recent time spans
  - The competition within the VNX type concerns probably rather genre, context or fashions than diachronic evolution

Results: do many NNs win in diachrony?

Table 2: Range of proportions of NNs divided in 6 periods of 20 years

Period → NN rate ↓	**1900-1919**	**1920-1939**	**1940-1959**	**1960-1979**	**1980-1999**	**2000-2019**
NPN dominant (NN rate under 1%)	66.3%	46.8%	42.2%	33.2%	16.2%	16.9%
NPN majority (NN rate 1%-30%)	24.6%	38.2%	42.5%	47.2%	58.2%	58.6%
Equipolent (NN rate 30%-70%)	4.4%	7.7%	7.5%	9.6%	13.2%	11.5%
NN majority (NN rate 70%-99%)	1.2%	2.6%	2.6%	3.3%	3.8%	4.3%
NN dominant (NN rate over 99%)	3.5%	4.7%	5.1%	6.8%	8.8%	8.8%
Total (types)	1360	1360	1360	1360	1360	1360

Results: families of NNs/NPNs

- Are there N1 or N2-based families (semi-specified constructions) for which the NN rate increases/decreases in diachrony?
  - Examples of members of **N1-based family**:
    - *trasporto passeggeri* ('passenger transport')
    - *trasporto merci* ('freight transport')
    - *trasporto rifiuti* ('waste transport')
  - Examples of members of **N2-based family**:
    - *trasporto merci* ('freight transport')
    - *vendita merci* ('goods sale')
    - *deposito merci* ('goods storage')
- Method: average rate of NNs for all the members of the family in the given period/year

Examples: N1-based families

- Increase: *assistenza* ('assistance') - to almost 80%, *trasporto* ('transport') - to almost 40%
- Decrease: *deposito* ('storage' or 'warehouse')

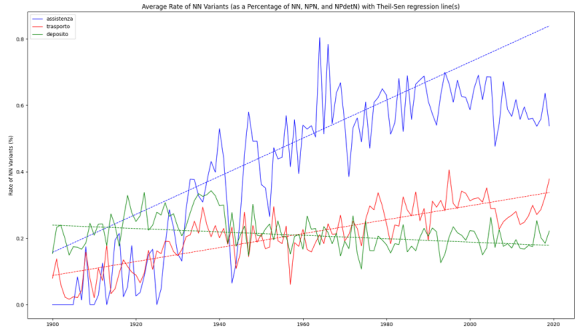

Figure 4: Rate of selected N1-based families over time

Examples: N2-based families

- Increase: *merci* ('goods'), *dati* ('data')
- Decrease: *viveri* ('provisions') – from 50% (1940's) to 10%

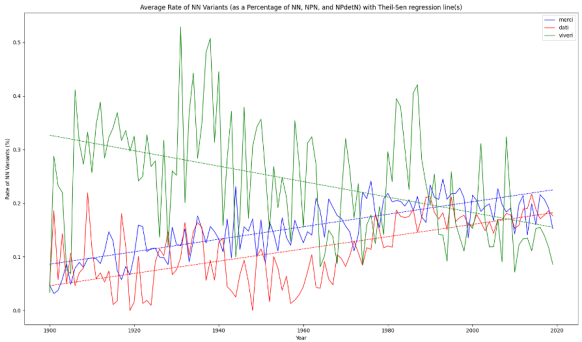

Figure 5: Rate of selected N2-based families over time

General trends: Clustering

- Based on the examples discussed so far, it seems that no clear *trends* can be detected.
- However, a functional clustering method developed by Dai et al., 2021 can be successfully applied to reveal clusters that exhibit similar diachronic behaviour.
- This clustering method was specifically chosen because it allows for a graphical interpretation of the resulting clusters through their central regions.
- The central region contains 50% of the cluster curves, representing the central development of the curves in each cluster over time, which facilitates easier interpretation.
- We conducted cluster analysis for various numbers of clusters and arrived at 8 such groups, since the corresponding average silhouette width—a measure of cluster compactness—was locally maximized for 8 clusters.
- The results were computed using the R package GET (Myllymäki & Mrkvička, 2023).

Clustering: 8 defined clusters

C1: Sharp increase during the period 1980–2000
C2: Sharp increase during the period 1960–1980
C3: Sharp increase during the period 1920–1940
C4: Sharp increase during the period 1940–1960
C5: High initial value followed by stagnation or slight decrease over time
C6: Gradual increase (slow)
C7: Gradual increase (very slow)
C8: Stagnation or extremely slow gradual increase

Cluster	No. of Competitors
1	44
2	41
3	40
4	28
5	53
6	367
7	399
8	388

Table 3: Cluster distribution of competitors

Clustering: visualization

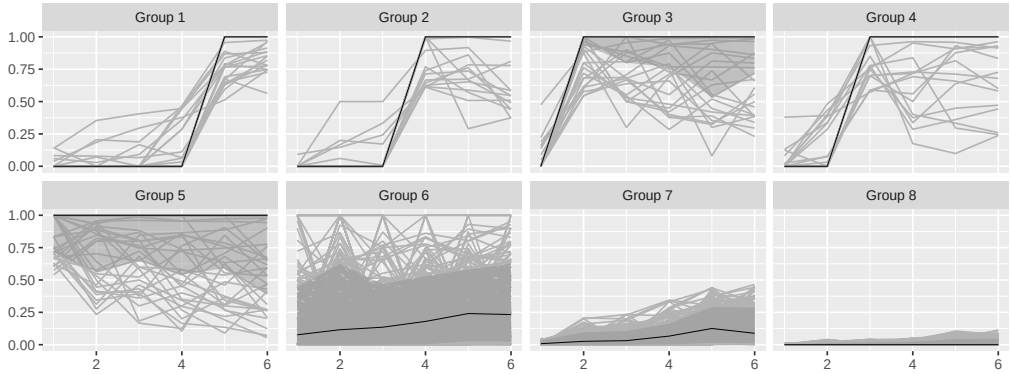

Figure 6: Visualization of the clusters

Conclusions

- The majority of competitors (58-85% of the curves, primarily in clusters 7-8, and some in cluster 6 with larger oscillations) show a modest upward trend.
- However, around 15% of the sample follows a different trajectory, characterized by a sharp increase that replaces the original NPN (if it was present).
  - This fashion trend emerges during various periods throughout the 20th century.
  - In general, most NNs and NPNs coexist in recent time spans with rather low proportions of NNs, which indicates that the competition within the VNX type is driven also by other important factors, such as genre, context or fashions

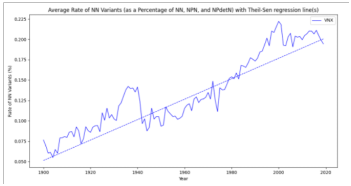

Figure 7: Average rate of NN variants over time

References I

Baroni, M., Guevara, E., & Pirrelli, V. (2009). Sulla tipologia dei composti n+n in italiano: Principi categoriali ed evidenza distribuzionale a confronto. In R. Benatti, G. Ferrari, & M. Mosca (Eds.), *Linguistica e modelli tecnologici di ricerca (atti del 40esimo congresso della società di linguistica italiana)* (pp. 73–95). Bulzoni.

Benveniste, É. (1966).Différentes formes de la composition nominale en français. *Bulletin de la Société de Linguistique de Paris*, *61*(1), 82–95.

Dai, W., Athanasiadis, S., & Mrkvička, T. (2021). A new functional clustering method with combined dissimilarity sources and graphical interpretation. In R. López-Ruiz (Ed.), *Computational statistics and applications*. IntechOpen. https://doi.org/10.5772/intechopen.100124

Goethem, K. V., & Amiot, D. (2019). Compounds and multi-word expressions in french. In B. Schlücker (Ed.), *Compounds and multi-word expressions* (pp. 127–152). De Gruyter. https://doi.org/doi:10.1515/9783110632446-005

References II

- Masini, F. (2009).Phrasal lexemes, compounds and phrases: A constructionist perspective. *Word Structure*, *2*(2), 254–271.
- Masini, F. (2019a). Competition between morphological words and multiword expressions. In F. Rainer, F. Gardani, W. U. Dressler, & H. C. Luschützky (Eds.), *Competition in inflection and word formation*. Springer.
- Masini, F. (2019b). Multi-word expressions and morphology [Last Accessed 5 August 2024]. In *Oxford research encyclopedia of linguistics*. Oxford University Press. https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-611
- Masini, F., & Scalise, S. (2012).Italian compounds. *Probus*, *24*(1), 61–91.
- Myllymäki, M., & Mrkvička, T. (2023). GET: Global envelopes in R. https://arxiv.org/abs/1911.06583
- Radimský, J. (2015). *Noun+Noun compounds in Italian: A corpus-based study*. Jihočeská univerzita. Edice Epistémé.

References III

Radimský, J. (2020).Are French NNs variants of N-PREP-N constructions? a corpus-based study of two competing patterns. *Linguistica Pragensia*, *30*(2), 156–186. https://doi.org/10.14712/18059635.2020.2.4

Rainer, F., Gardani, F., Dressler, W. U., & Luschützky, H. C. (Eds.). (2019). *Competition in inflection and word formation*. Springer.

Thornton, A. M. (2019). Overabundance: A canonical typology. In F. Rainer, F. Gardani, W. U. Dressler, & H. C. Luschützky (Eds.), *Competition in inflection and word-formation* (pp. 223–258, Vol. 5). Springer International Publishing. https://doi.org/10.1007/978-3-030-02550-2_9

Tollemache, F. (1945). *Le parole composte nella lingua italiana*. Edizioni Rores di Nicola Ruffolo.