Glossa a journal of general linguistics

Garzonio, Jacopo and Cecilia Poletto. 2018. Exploiting microvariation: How to make the best of your incomplete data. *Glossa: a journal of general linguistics* 3(1): 112. 1–21, DOI: https://doi.org/10.5334/gjgl.556

RESEARCH

Exploiting microvariation: How to make the best of your incomplete data

Jacopo Garzonio and Cecilia Poletto

Department of Linguistic and Literary Studies, University of Padova, Piazzetta Folena 1, 35137 Padua, IT

Corresponding author: Jacopo Garzonio (jacopo.garzonio@unipd.it)

In this article we discuss the use of big corpuses or databases as a first step for qualitative analysis of linguistic data. We concentrate on ASIt, the Syntactic Atlas of Italy, and take into consideration the different types of dialectal data that can be collected from similar corpora and databases. We analyze all the methodological problems derived from the necessary compromise between the strict requirements imposed by a scientific inquiry and the management of big amounts of data. As a possible solution, we propose that the type of variation is *per se* a tool to derive meaningful generalizations. To implement this idea, we examine three different types of variation patterns that can be used in the study of morpho-syntax: the geographical distribution of properties (and their total or partial overlapping, or complementary distribution), the so-called *leopard spots* variation, and the lexical variation index, which can be used to determine the internal complexity of functional items.

Keywords: linguistic atlas; Italian dialects; microvariation; morpho-syntax

1 Introduction

In this paper we discuss different distributional patterns that emerge taking into consideration linguistic, and in particular syntactic atlases and databases. We will base our investigation on the ASIt database (Atlante sintattico d'Italia), the Syntactic Atlas of Italy (cf. Benincà and Poletto 2007; Di Nunzio, Garzonio & Pescarini 2014),[1] one of the first tools of this type, and on further data that have emerged from our enterprise gathering data for ASIt but have not been published on the web yet. We will start by showing that big linguistic enterprises, like databases, atlases like the ASIt and all sorts of corpora always contain a certain amount of "noise". They are by definition always incomplete when the hypothesis we want to test is very detailed. Since this is inevitable, we will not present statistical methods to circumvent this problem but consider some alternative strategies that can help us to find interesting theoretical clues even in data that provide by definition a coarse-grained picture of the linguistic reality. If it is true that big amounts

[1] The ASIt project is a longstanding enterprise which started in the 1990s with the aim of gathering syntactic annotated data on the dialects spoken in Northern Italy and later also in Central and Southern Italy. The original method was a survey of written questionnaires sent to the speakers since the project did not have any funding. From the early 2000s, the questionnaires have been administered directly by collaborators of the project and the tagging system has been completely reorganized. The data collection started in the Northern area and was then expanded to the South with a series of new questionnaires adapted to the properties of the Southern varieties but with the same translation methodology. It continues at present with a set of questionnaires that are visible on the website, and can be consulted at http://asit.maldura.unipd.it/, to which we refer for a set of the phenomena investigated, and more details for the present methodology of gathering the data. Not all data are published on the web and are first ordered, tagged and analyzed by the research group. At present (April 2018) the data set includes tagged sentences from 409 Speakers representative of 289 inquiry points for a total of 77.210 tagged items, i.e. sentences (2.263.924 total words).

of data are never precise enough for a very detailed hypothesis, we can still try to exploit the peculiarity of a blurred image to single out the general outlines of the linguistic panorama, which would remain otherwise uncovered. The type of "big data" we can have in dialectology cannot rival with big data in other disciplines, but not even with big data in other linguistic domains. Since dialects are micro-variants minimally different from each other, it is clear that we cannot have the same amount of linguistic information we find for instance in typological enterprises that cover a broad spectrum of languages. However, we believe that they can nonetheless be interesting. In this way, using big amounts of data mining can nicely complement our introspective type of empirical evidence. In other words, the reason why big amounts of data are always "noisy" is that we ask them too much, and the questions we ask are not apt to the type of evidence we have. The solution to the problem we will present is the following: up to now we have only used corpora to look at the presence versus absence of a given structure or phenomenon in a given language or variety and relate it to other structures. An innovative way to think about big data sets and tailor our questions on the linguistic evidence provided by big amounts of data is to consider the type of variation itself as a clue indicating different natural classes of linguistic phenomena. We will show that there are various ways to make use of incomplete and noisy data and that there are advantages in having a blurred but general picture. The gist of the work is in a way similar to the way typological work provides broad overviews of phenomena and highlights overall tendencies, although what we do is restricted to single homogenous linguistic areas. In a sense, this work constitutes the first steps towards a theory that uses distributional patterns as evidence to exclude or favor a family of analyses with respect to other logical possible ones.

In section 1.1 we discuss several problems in dealing with dialectal big amounts of data. In section 2 we discuss different types of distribution and analyze some cases of generalizations and theoretical proposals based on these different distributions: more precisely, in 2.1 we discuss different patterns in the geographic distribution of morpho-syntactic properties; in 2.2 we discuss *leopard spots* variation; in 2.3 we discuss the lexical variation index. Section 3 contains some conclusive considerations.

1.1 Some notes on the databases used in this article

Any empirical investigation requires experiments of some sort, i.e. data gathered in a controlled environment for the variable that we are testing, which are replicable for the same set and comparable with other sets coming from different linguistic environments. These three properties of reliable data: a) controlled b) replicable and c) comparable apply to whatever type of linguistic variety and type of population (adults, children, normal or impaired) we are working on. The recent rise of experimental linguistics has brought the construction of experiments in the focus of research, since no generalization is possible without a founded pool of data. In the ideal case, also the investigation of non-standard or minority languages should follow the same rules we find in language acquisition, studies on language impairment of different sorts, neurolinguistics and all the other empirical fields to which experimental linguistics has recently opened itself. Replicability is the chief property of empirical work: if we cannot replicate our data, they cannot be the object of a scientific investigation. In order to achieve replicability, we need to control the setting of our test. Otherwise, unforeseen factors may intervene and blur the picture of the phenomenon we intend to investigate. On the other hand, these absolutely fundamental requirements clash against the reality of the complex distribution of non-homogeneous linguistic situations, i.e. linguistic environments where more than one language share the stylistic spectrum speakers have at their disposal and where the speakers can modulate their linguistic behavior in relation to the situation they find themselves in. The particular

situation of non-standard languages, which are often spoken only in familiar contexts, and have either lost, or never had a standardized version, i.e. higher levels of formalized or written style, with an established normative tradition, poses challenges that the investigation of standard languages does not have to meet. From this point of view, the situation in Italy is particularly problematic because non-standard languages (i.e. the so-called Italian dialects) are not local varieties derived from the standard Italian language, since they independently derive from Latin (so standard Italian is an Italian dialect in this sense). At the same time they are all closely related languages. Furthermore, the type of investigation we can carry out, the tasks we use, the type of informants, in essence all the variables of our experiment depend on the type of sociolinguistic situation we are dealing with. There are cases in which the non-standard language is not stigmatized from the environment, as it is perceived as a sign of identity which does not pose any problem to the investigators. There are also cases in which speakers control themselves in order not to produce the phenomena of the language variety we are interested in because they are perceived negatively, as the phenomenon of stigmatization attests (see among others Cornips and de Rooij 2014).

We will mainly use data from the Northern Italian dialects because this area has already been established in traditional dialectological work as a homogeneous area from the middle ages until today, where we still see common trends of change, which are not present either in the Central and Southern dialects or in standard Italian (Poletto 2012). As an instrument, we will use the ASIt database, which contains syntactic and morphosyntactic data from over 280 Italo-Romance varieties, collected as questionnaire translations and searchable through a large set of syntactic and morpho-syntactic tags. It corresponds to the modern version of the classic AIS (*Atlante Italo-Svizzero*, i. e. the *Atlas Italiens und der Südschweiz*; Jaberg and Jud 1928–1940), the data of which were collected 100 years ago. The AIS is also available as searchable maps on the web.[2] Enterprises of this type can be considered as a compromise between a narrow aimed qualitative inquiry, which requires a set of narrow targeted experiments carried out within a single language with a statistically relevant number of informants, or even a case study of a single speaker, and real big amounts of data acquired through the many websites where different dialects are spoken nowadays, which are completely uncontrolled both in terms of the speakers and in terms of data, which are taken from free speech and not through a controlled experiment. Case studies of a single speaker are perfectly justified in dialectology as they are in neuropsychological and acquisition studies for various populations of speakers, since the study of dialectology is precisely about micro-variation and if the study is detailed enough, one finds single individual grammars (cf. Ludlow 2011: 44–46, on idiolects in formal approaches). So, instead of looking for the common features shared within a certain population, dialectological enterprises turn upside down the focus of the research and seek out the property of the grammar of a single individual much as a single DNA can be mapped.

Traditional dialectology has always used translation tasks performed by a single speaker per dialect as the most manageable type of test, capitalizing on the idea that speakers are aware of when they are talking their dialect and when they are talking the standard language. This state of affairs is becoming more and more difficult on the view that speakers are nowadays all bilinguals with the standard language (which they were not a hundred years ago not only in Italy), with different dominant variants, mixed variants, massive code switching, passive or evanescent speakers etc. (Berruto 1987; Dorian 1989; Dal Negro 2004). So, the choice of speakers which was knowingly made by traditional

[2] The searchable version of the AIS, developed by the *Consiglio Nazionale delle Ricerche*, is available at the site http://www3.pd.istc.cnr.it/navigais/.

dialectologists might not be enough nowadays and present day investigations probably contain more noise than they might have contained a hundred years ago. On the other hand, various dialectological domains have witnessed in recent years a tendency towards the revitalization of the local varieties, but it is also reported that there is a tendency towards *koinization*, i.e. the creation of a regional variety which is understood in a bigger geographical area. The "koine" variety displays features of the more prestigious varieties of the area (for instance the variety of Turin for Piedmontese; Cerruti and Regis 2014). As a consequence, single properties of a local dialect have spread or disappeared, depending on its sociolinguistic status and on other sociolinguistic factors, like the desire to be considered as part of a community, or to set a group apart from others, etc. However, we still believe that for a big enterprise of the type required by a syntactic atlas, the way to begin testing unknown ground can still be translation tasks as are used in the AIS and ASIt. However, they cannot be the last step of our inquiry, but only constitute the first probing procedure to determine the phenomena that might be interesting out there. The necessity to use on the one hand translation tasks and on the other to limit the amount of speakers of a single dialect to a number which often does not range above 5 also has a purely organizational and resource management reason. If we want to gather data from many dialects, the task must be manageable, and we cannot afford to have too many speakers per dialect, otherwise the net of the inquiry points will be too broad and we will never be done. It would be advisable to find other ways to gather data from the web, but at present we have not yet developed the necessary techniques to clean the data enough for them to be interesting from a microvariation perspective. It is not possible to use massive statistical procedures on data that are by definition restricted by the local variant, unless we content ourselves in looking at macro-regional varieties, which is interesting, but in a sense a less powerful microscope to investigate micro-variation.

Therefore, databases of non-standard varieties are still by necessity a compromise between the strict requirements imposed by a scientific inquiry and the management of big amounts of data, which cannot be tagged fully automatically, but require complete or at least partial manual tagging. This cautionary note has to be kept in mind when we start generalizing over maps and distributions of phenomena we will investigate here.

2 How to use distributional patterns to enhance our understanding of variation

In order to show that geographical distributional data are per se interesting to a microvariation perspective, in this article we will single out three possible distributional types and determine to which type of phenomenon each type of variation is related.

We will first take into account the "classic" method of comparing the geographical distribution of different phenomena and consider the theoretical import of different distributional patterns, in particular those in which two phenomena a) completely or b) partially overlap, or c) are in complementary distribution on a map.

Then we will consider a type of distribution which can meaningfully be exploited only when looking at a dialectological area, and not to typologically different languages and which can provide us with interesting observations that we would not be able to see on the basis of a different type of investigation. The term we will adopt is the one commonly used by Romance dialectologists, namely *leopard spots*. With this definition what is meant is that a given phenomenon is found scattered in a whole area not in a homogeneous way but occurs precisely with an apparently random distribution. So, dialects that are close to each other might differ with respect to the phenomenon under study, which might however be observed somewhere else in the geographical domain under investigation. Since traditional dialectologists (see for instance the very insightful work by Bartoli 1945) have

already seen the potential of micro-variation to understand the principles of grammar and linguistic change, they had already noticed tendencies that would be explained later on by sociolinguistics. For instance, when a given phenomenon is only found in the center of an area, it is an innovation or when it is found at the borders it is most probably a remnant of a phenomenon that had a much larger distribution in the whole area. Leopard spots phenomena are interesting under this perspective precisely because they cannot be made sense of within such a framework. In other words, leopard spots phenomena are the odd man out with respect to geographical distribution, and in our view are so, because leopard spots is the distributional pattern corresponding to multifactorial phenomena that emerge only when several necessary but not sufficient conditions cluster to allow the development of the property. This means that we can identify multifactorial phenomena simply by looking at their distribution.

The third type of distributional pattern we will look at is based on the amount of etymological variants and has *prima facie* little to do with syntax and consists in considering the lexical amount of variants for a given functional item. We will see that within the same paradigm, take for instance quantifiers, or wh-items or modal verbs, there are some that are etymologically extremely stable in a linguistic domain while others vary a lot. We will develop an idea put forth in Poletto (2012), according to which the lexical variation index is a function of how semantically and syntactically complex a (functional) element is. This type of method is clearly not only syntactic and requires once again a rather stable area of inquiry, where we can filter out phonological variants and sort out whether there are real different etymological stems at play.

2.1 Overlap or complementary distribution and what they might mean to a syntactician

The ASIt is particularly useful in order to capture these types of overlaps because the data were collected with the precise idea of delimiting the areal distribution of many different phenomena and syntactic constructions and investigate their interaction. This is a rather classic methodology that has been used in work on micro-variation and helps us to determine whether apparently unrelated phenomena are actually related. We will provide some examples of these types of distribution using the ASIt data. The first case we consider is the one of coincidence, i.e. when two phenomena completely overlap.

A. A well-studied case of a *complete overlap/coincidence* of phenomena is the strict relation between preverbal negation and the impossibility to negate true imperative verb forms (see Zanuttini 1997: 105–107):

(1) *Standard Italian*

a. *Non mangialo!
NEG eat.IMP.2SG = it

Redondesco, Mantua

b. Mángel mia!
eat.IMP.2SG = it NEG
'Do not eat it!'

Zanuttini (1997) formulates the following generalization: when we have a true imperative form, i.e. a form which is morphologically unambiguously only imperative, the preverbal negative marker is not compatible with it. This means that the verbal form changes into a suppletive one, which varies according to the dialect (it is the infinitive in standard Italian and in many other cases). Zanuttini reports dialects that have an infinitival form, a gerund or a subjunctive, but they all obey the same rule, i.e. when the preverbal negative marker is present, the verbal form changes. On the other hand, when the negative marker is

postverbal, the true imperative remains also in negative imperatives. The two phenomena that overlap here are a) the position of the negative marker and b) the morphological form of the imperative verb. To show that they are related has brought Zanuttini to an interesting analysis according to the following lines: the preverbal negative marker requires TP and true imperative forms lack this projection. Therefore, true imperative forms are only compatible with postverbal negative markers that do not require TP. The point here is not whether this analysis is correct or not, but simply to show the power of such an empirical observation based on the ASIt and AIS databases. Since 1997 we have found a number of dialects that disobey this generalization in both senses, i.e. there are some dialects where the imperative form changes although the negative marker is postverbal, as noted for some Emilian varieties in Benincá and Poletto (2005):

(2) *Albinea, Reggio Emilia* (Benincà and Poletto 2005: 245)
Movrat mia!
move.INF = yourself NEG
'Don't move!'

This has been explained by Benincà and Poletto by assuming that these dialects still have a null preverbal negative marker, since they are close to dialects that still present a discontinuous form of negation similar to French *ne…pas*. However, there are also dialects where the preverbal negative marker seems to tolerate a true imperative form, like the variety of Cortina d'Ampezzo (at least the one of older speakers) taken into consideration by Vai (1998).

(3) *Cortina d'Ampezzo, Belluno* (Vai 1998: 660)
No laóra adès!
NEG work.IMP now
'Don't work now!'

These data are rather sparse, but here one might have to do with a different negative marker: notice that the preverbal negative marker is homophonous with the pro-sentence negative particle. That here we have to do with a different type of preverbal negative marker is attested by Venetan varieties, where the two are phonologically distinct, since the standard preverbal negative marker is pronounced [no] with a closed vowel, while the pro-sentence negation is pronounced [nɔ] with an open vowel. In negative imperatives the negative marker used is the one with the open vowel, i.e. pro sentence negation not the one with the closed vowel, the regular sentential negative marker.[3] It remains to be seen why pro-sentence negation tolerates a true imperative form, but in any case we can conclude that Zanuttini's generalization is not disconfirmed by these data.

[3] The pattern of Venetan dialects is rather complex, since there is an alternative strategy, already noted by Kayne (1992), namely the possibility to insert a modal auxiliary in the imperative form, which then tolerates the standard negative marker and cannot be used in the positive variant.

(i) a. *Paduan* (Kayne 1992: ex. 17)
No stá parlare!
NEG AUX.IMP talk.INF
'Don't talk!

b. *Paduan* (Kayne 1992: ex. 18)
*Stá parlare!
AUX.IMP talk.INF
'Talk!'

See Oliviéri and Poletto (2018) for a discussion and an alternative proposal to the one of Zanuttini, which is not discussed here because the focus of this work is not strictly theoretical but methodological.

An important methodological point in looking at geographical distributional patterns is the fact that any generalization can be violated by a small number of dialects. But if this is the case, one should look a bit more closely into the few dialects that do so, because there might be another explanation at hand which allows us not to throw away precious observations confirmed by hundreds of dialects simply on the basis of one dialect. On the other hand, we should not simply set aside data coming from exceptional dialects, because they might be decisive for the analysis. So, on the one hand, percentages have in principle nothing to do with this type of investigation, on the other hand, when a generalization is very robust, it is advisable to double check and not stop when one dialect seems to behave differently.

B. The second case of a typical geographical pattern which provides us with insights into the principles of variation is the case of *inclusion*. We exemplify it here with another well-known phenomenon, namely the doubly filled comp filter violation found a bit scattered in many Italian dialects. The so-called wh + COMP construction was already mentioned in Poletto and Vanelli (1995), where the following generalization is put forth: the varieties that present a complementizer in independent interrogative clauses also present it in embedded questions, but not all the varieties with wh + COMP in embedded questions also display it in independent ones.

(4) *Aldeno, Trento*

a. Dime sa che la magna la Maria.
tell = me what that SCL = eats the Mary
'Tell me what Mary is eating.'

b. Sa (*che) fé?
what that do.2PL
'What are you doing?'

This pattern of inclusion can also be described on the basis of a one way relation of the type *if A then B*. In our case, if a dialect has the structure wh + COMP in main clauses, then it also has it in embedded clauses. This pattern is the dialectological counterpart of the implicational relations that are also much used in typological work. The distribution of wh + COMP can be represented as in Figure 1, based on the ASIt translations for the Italian sentences *Dimmi dove è andato Giorgio* 'Tell me where Giorgio has gone' and *Dove è andato?* 'Where is he gone?'. The data show that many varieties present the complementizer only in the embedded question, while some others have it in both the embedded and the main question. Only in one case have we found the opposite pattern.[4]

Implicational relations and often even implicational scales can be seen as generalizations of a higher level that point us towards a complex explanation. On the other hand, in a dialectological area they can also be the expression of a diachronic development, which in the case of our example goes from embedded clauses to main clauses. Notice however that saying that this is only the reflex of a diachronic path is not an explanation but is just reformulating the question as to why the development always goes in this direction and never the opposite. Furthermore, this type of inclusion phenomena might also include more than one step. Parry (2003) has shown that in Piedmontese the wh + COMP structure has developed out of relative clauses, but once again, this points towards the direction that between embedded interrogatives and relative clauses there could be a relation that has not yet been fully explored. The cartographic approach has stopped at the observation that relative wh-items and interrogative wh-items target different projections, since they

[4] The divergent variety is the Ligurian dialect spoken in Arcola (La Spezia), at the boundary between the Gallo-Italic and the Tuscan macro-areas.

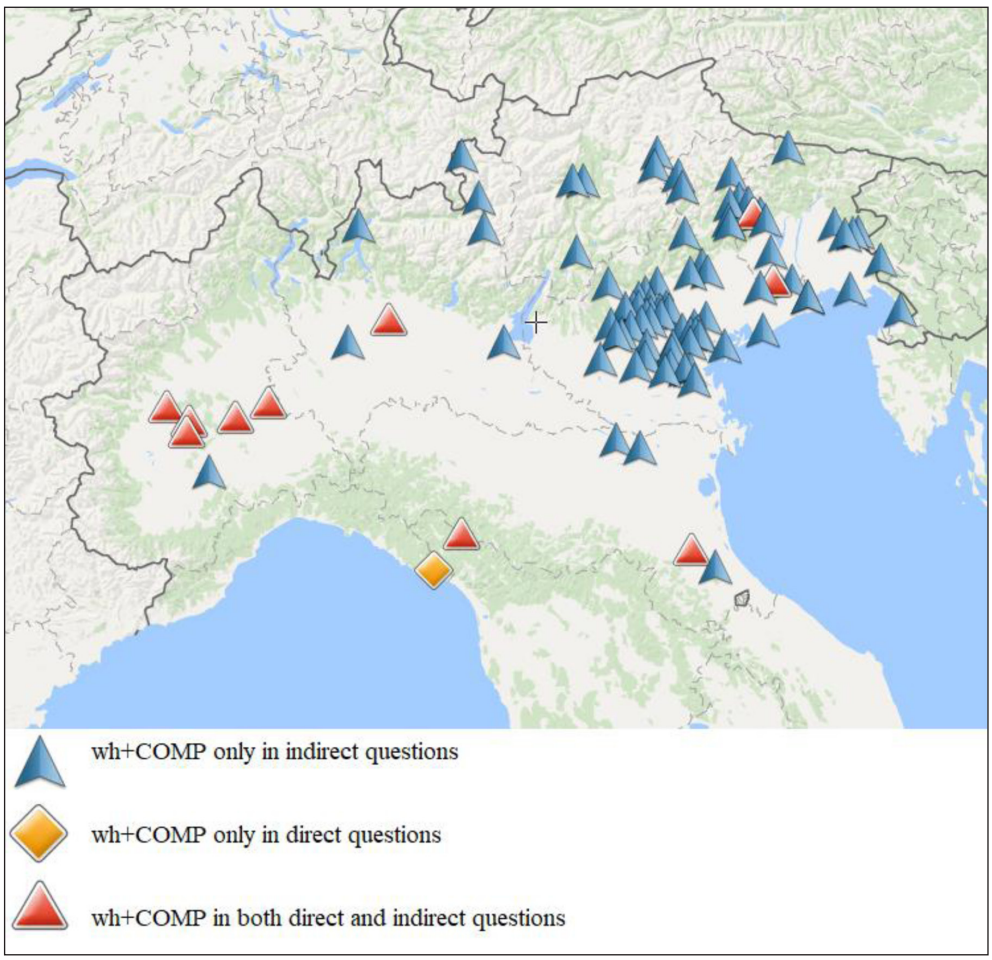

Figure 1: The distribution of wh+COMP in Northern Italian dialects.

occur at different side of left dislocated elements. However, the very fact that in several languages, including Romance, wh-items can cover both relative and interrogative functions points to the direction that there must be a closer connection. Without entering a theoretical discussion on relative and interrogative clauses, we point out that implications like the one above can constitute the basis to look beyond the received view on a given phenomenon. If one wants to look more closely, besides dialects like those of the Veneto area, where all wh-items are obligatorily followed by a complementizer in embedded clauses, there are others, like dialects spoken in Trentino, where only some wh-items tolerate the complementizer. Since these wh-items are the ones that can occur in free relative clauses (usually the items corresponding to 'who' and 'where'; cf. Garzonio 2007), once again we see that the connection between embedded interrogative clauses and relative clauses has to be taken into account to explain the distribution of the wh + COMP construction. Looking even closer, we can see that in the Friulian area several varieties systematically tolerate this construction in standard main interrogatives only with *who*, so that we have a structure *who-that* which is similar to the one also reported in Quebec dialects and some conservative French dialects.

(5) *Barcis, Pordenone*
a. Cu c' al vai là via?
who that SCL = cries over there
'Who cries over there?'

b. Cu c' al mangia le patate?
who that SCL= eats the potatoes
'Who is eating the potatoes?'

c. Cu chi i no invita?
who that SCL= NEG invites
'Who won't they invite?'

While this construction is possible both with subject and object *who*, it seems that in some varieties it is preferred with subject *who*, while with object *who* it is optional (cf. (5) and (6)):

(6) *Barcis, Pordenone*
Cui aj o dismendead?
who have =SCL forgotten
'Who have I forgotten?'

This observation will then have to be treated together with several facts that put wh-movement of the subject aside from wh-movement of the direct object or even of other PP objects and adverbials. This is not the place to discuss which analysis might account for this, but a set of generalizations as the following one clearly points towards a connection between relatives and embedded interrogatives:

(7) a. If wh + COMP occurs in main interrogatives, then it occurs in embedded interrogatives.
b. If wh + COMP does not occur with all wh-items, it occurs first with those that are also possible in free relative clauses.
c. If only one wh-item allows for the wh + COMP construction in main interrogatives, it is the one that corresponds to the subject.

Notice that this set of generalizations helps us to distinguish what is happening in the Northern Italian Dialects with respect to a phenomenon that might look very similar, namely the occurrence of the complementizer *dass* 'that' after wh-items in Bavarian (see Bayer and Brandner 2008). In Bavarian the complementizer occurs first with complex wh-items and not with wh-words and is generally never found in main interrogatives. So, although the phenomenon looks alike, the pattern of the distribution is rather different.

C. The third case of geographic pattern which can be of interest to the theoretical linguist is the one of complementary distribution. It could definitely be a chance phenomenon, on a par with coincidence and inclusion, but could also be interesting, especially if we find it reproduced in different dialectal areas. A case of observable *complementary distribution* is given by interrogative particles derived from the wh-system: only varieties without enclisis of subject clitics in questions (or without subject clitics at all) have developed yes/no question particles from the wh-item corresponding to 'what' (see also Lusini 2013). Florentine has subject clitics (*tu* in (8a–b–c)), but lacks enclisis in interrogatives (compare (8b) vs (8c)). However, like other Central and Southern Italian varieties, it has a yes/no interrogative particle identical to 'what' (8b).

(8) *Florentine, Florence*
a. Maria, che tu conosci anche te, è a Napoli.
Mary that SCL= know.2SG also you is at Naples
'Mary, who you too know, is in Naples.'

b. **Che** la compri, o un tu lla compri?
what it = buy.2SG or NEG SCL = it = buy.2 SG
'Are you going to buy it or not?'

c. *La compri-tu?
it buy.2SG = SCL
'Are you buying it?'

This distribution is represented in Figure 2, which is based on the ASIt translations for the Italian sentences *Parti subito?* 'Are you leaving immediately?' and *Hai visto un bambino con i capelli rossi* 'Have you seen a red-haired boy?'.[5] The areas with clitic subjects and interrogative particles identical to *what* are complementary, with Tuscan varieties like the one of Florence at the boundary. Crucially, all Tuscan varieties with subject clitics have

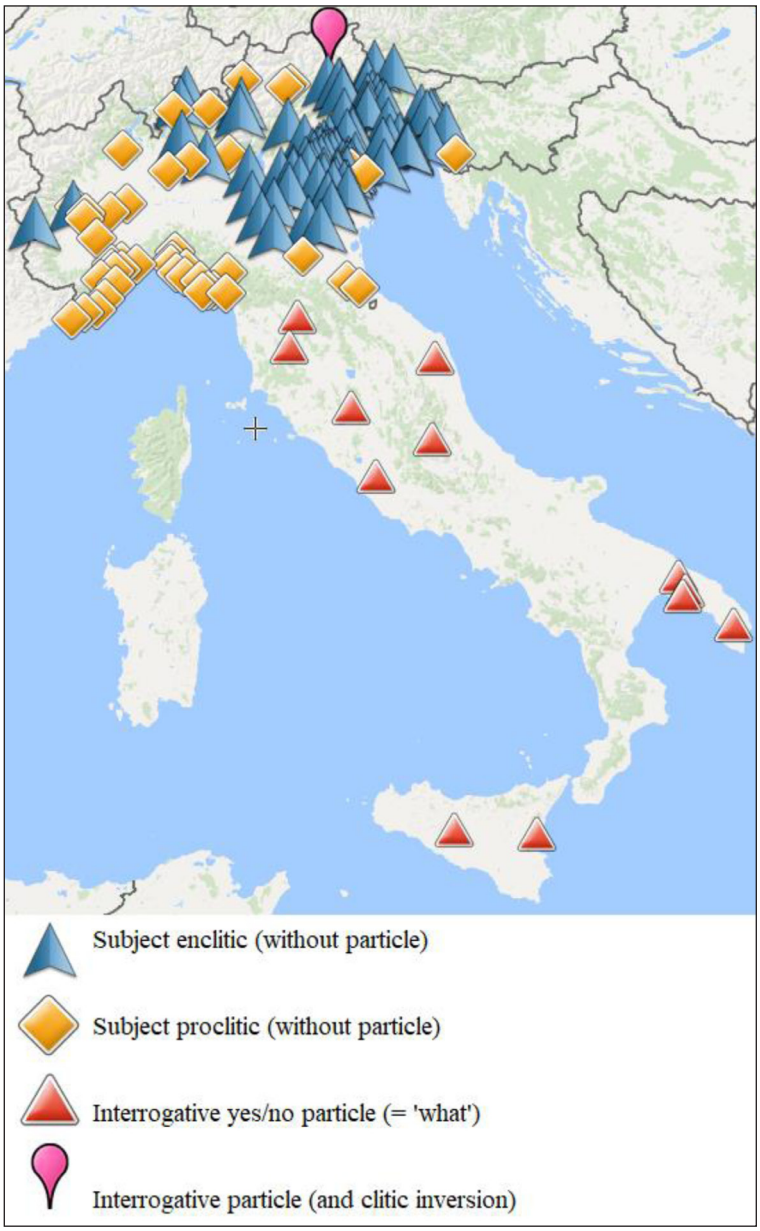

Figure 2: The distribution of interrogative clitic inversion and interrogative particles in Italo-Romance.

[5] We have selected two different questions with a second singular subject, since the ASIt has different questionnaires for Northern and Southern varieties.

lost enclisis, i.e. clitic inversion, in questions. Notice that some Ladin varieties have both subject clitics (with enclisis) and an obligatory interrogative particle, but the latter is not related to the wh-item system, since it derives from the Latin temporal/aspectual adverb *post*. We provide here an example from the variety of Selva di Val Gardena (where the particle is reduced to *a*):

(9) *Selva di Val Gardena, Bolzano*
Pëies' **a** riësc vía?
leave.2SG PRT immediately
'Are you leaving immediately?'

Concluding, this type of methodology has been used by traditional dialectologists and is still used today in formal frameworks and can be only be adopted when we are comparing two phenomena and trying to establish whether they are related or not. Coincidence would then be interpreted in a way to make the two phenomena depend on the same abstract property. Inclusion would be interpreted in a way such that the phenomenon that is more largely represented is a necessary but not sufficient condition for the occurrence of the second. Complementary distribution would be a case of alternative checking (Obenauer 2004) of the same property, so that you can only have either the first or the second phenomenon. Still, the distribution we find could only be by chance, but if we have enough varieties, the probability that we only have to do with chance reduces the bigger our sample is. This means that it is possible and advisable to use big amounts of data, since they are not "noisier" than little/qualitative data.

2.2 Leopard spots variation

Typical cases of leopard spots variation are rather common in dialectal areas and correspond to structures which distribute over the entire domain but are not grammaticalized in all dialects. One might wonder whether the so-called leopard spots distribution is due to some warp effect in the data, for instance to the fact that data are missing from some varieties but present for others. That this cannot be the case is attested by the fact that this type of distribution is extremely common even in well investigated areas, as was already known by traditional dialectologists.[6] Furthermore, in several cases, the distribution is clearly not due to a gap in the data, but we know for sure that the property is distributed in an uneven way through a given territory. A rather clear example of leopard spots variation is the occurrence of partitive objects (that is nominal expressions introduced by the preposition corresponding to 'of', from now on NPOs') under negation, similarly to pseudo-partitives under quantifiers and negation of standard French (see among many others Stark 2008; Luraghi 2012).[7] This phenomenon is attested in some, though not all Ligurian and Piedmontese varieties, in Alpine Lombard and in the eastern part of the Emilian area (see Garzonio and Poletto 2017). The presence of the preposition corresponding to 'of' (in some dialects conflated with the definite article) is triggered only by

[6] For instance, already Ascoli (1873) pointed out that the properties he considered as typical of "Ladin" varieties (i.e. Rhaeto-Romance dialects), like the palatalization of velars before /a/ or the sygmatic plural, are not attested in all the dialects but can freely emerge in a scattered pattern among them.

[7] Another possible example is the distribution of a set of interrogative structures in the Northern Italian domain (which we are dealing with here, but the same type of observation can be made for Gallo-Romance varieties), where we see that dialects can either use one or more of the following structures to express main interrogatives: a) obligatory clefts; b) wh-COMP (see above); c) wh-in situ; d) wh-doubling. The distribution of these four structures definitely requires a multifactorial analysis, since a single dialect can allow for one, two, three or even all these structures in addition to subject clitic inversion. Since this is very complex and would require a whole article by itself, we have discussed a more delimited case as the one of partitives under negation.

sentential negation, and the nominal expression is always interpreted as indefinite, like in the case of standard partitive articles:[8]

(10) a. *Cicagna, Genua*
Nu t'acati mai **de** meie.
NEG SCL = buy never of apples

b. *Sondalo, Sondrio*
Te'n crompesc mai **de** póm.
SCL = NEG buy never of apples

c. *Cesena*
T'an cumpar mai **dal** meili.
SCL = NEG buy never of.the apples
'You never buy apples.'

NPOs are found in a sporadic way in various Gallo-Italic dialects so that the phenomenon is scattered across the whole area. In Figure 3 we provide the distribution of the presence of *of* under negation in the translations of the Italian sentence *Non compri mai mele* 'You never buy apples'.

Figure 3 shows that NPOs are rather common in the Ligurian area, but by no means confined to it, and can be found in all Northern dialectal areas, i.e. Piedmont, Lombardy, Emilian, but not in the Veneto and Friulian areas.[9] This means that, although the Ligurian area evidently has the highest concentration of the phenomenon, the cluster of properties required for it to occur are by no means unique to this area. This does not yet point us towards an analysis but provides us with a procedure we can use to establish exactly which morpho-syntactic properties are necessary for the phenomenon to manifest itself. If we systematically compare the dialects that have the property with those that do not, we should be able to find out which necessary conditions are at play here. The fact that

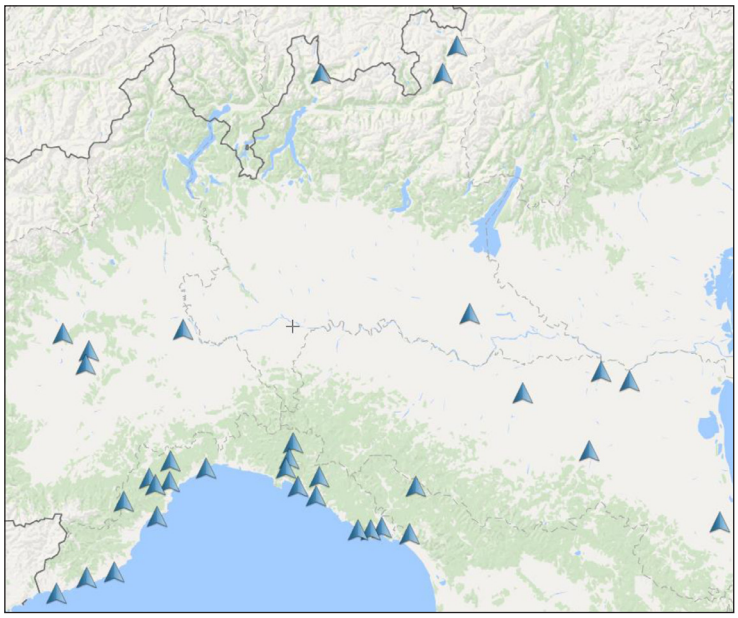

Figure 3: Presence of a partitive object under negation.

[8] Notice that this phenomenon can be observed also from the point of view of the *inclusion* pattern examined in the previous section, since varieties with negative partitive objects are a sub-set of the varieties that have partitive articles, although the morphology of the two forms is different, since in Italo-Romance partitive articles include the definite article while NPOs only in some varieties.

[9] Cases of pseudopartitives under the wh-item corresponding to 'how many', etymologically the same item meaning 'many', are attested in the Ladin area spoken in the Dolomites.

the whole North-Eastern area has been left out is suggestive of the fact that this is also the only area where postverbal negation has not advanced much further than in Northern colloquial Italian, which uses the postverbal negative marker *mica* only in presuppositional and implicature contexts, as already pointed out by Cinque (1976). Therefore, it might be the case that one necessary condition for partitives under negation to arise is the type of negative marker, which must be of the postverbal type. However, if we look at the Ligurian area, where partitives are very common, the type of negation is also broadly a preverbal one. This means that a theoretical explanation has to go into the direction of looking into the different types of preverbal negation. Another possible necessary condition for partitives under negation to occur might have to do which the licensing of bare nouns with stage level as well as with individual level predicates, which are attested in some dialects, while in others require a definite expletive. In Garzonio and Poletto (2017) we have proposed that these structures contain a null quantificational noun, whose modifier is the negative marker, which is etymologically a minimizer. As shown in (11) for French, the partitive marker *de* is actually a real case marker and not simply an indefinite determiner as partitive articles (Cardinaletti and Giusti 2016). We argue that the variation observed in Figure 3 is based on three main factors: a) the (im)possibility to license the null noun, b) the type of null noun licensed (either AMOUNT or NUMBER or both), and c) the categorial status of the item *of*.

(11) a. $[_{\text{NegP}}$ *pas* … $[_{\text{QP}}$ [~~*pas*~~] $[_{\text{Q}^\circ}$ AMOUNT $[_{\text{KP}}$ *de* [‘whole’]]]]]
b. $[_{\text{NegP}}$ *pas* … $[_{\text{QP}}$ [~~*pas*~~] $[_{\text{Q}^\circ}$ NUMBER $[_{\text{KP}}$ *de* [‘set’]]]]]

We have proposed that the null noun is normally licensed by the postverbal negator, a modifier of nominal nature (see Manzini and Savoia 2011 on this). In order to derive the pattern represented in Figure 3 correctly, we have to assume that in Ligurian varieties the postverbal negator is null and preverbal negation is sufficient to license the whole structure (it can be argued that in this case the preverbal negation corresponds to the negative operator). This evidently requires further investigation on the properties of Ligurian negative markers. The reason why we propose that two types of null nouns are involved (AMOUNT vs. NUMBER in (11)) is linked to the fact that NPOs can be obligatory only with plurals (this means that in such varieties the null noun can be only NUMBER). Finally, a further problem to investigate to explain Figure 3 is that Friulian and some Venetian varieties have a complex item corresponding to ‘of’ (*da*, formed by two separate functional prepositions, *d(e)* ‘of’ plus *a* ‘to’), which apparently cannot encode partition in these structures. The distribution of the different case markers is thus most probably related to the occurrence of NPOs.

We are not going to provide here a detailed analysis of all the conditions necessary for this phenomenon to occur, since this is not the focus of the present article. We only point out that the leopard spots distribution of some phenomena directly points towards a multifactorial analysis of the phenomenon under consideration. Furthermore, if there is any area spared from the spots, as it is the case here, we should look precisely what this area lacks because it might be one of the preconditions for the phenomenon to establish itself. We conclude that the case of the leopard spot pattern, although it has not been used so far in theoretical analyses on microvariation data, is probably the most precise distributional tool to assess which syntactic properties might play a part in the occurrence of a given phenomenon.

2.3 A new investigation strategy: The lexical variation index (LVI)

When handling big amounts of data, we should not make the conceptual mistake of trying to get the same type of highly refined but statistically non-significant evidence we get from qualitative work. We should rather invent new methods to “sieve” interesting

observations from the mass of the data we have at our disposal. Therefore, big amounts of data are not simply to be treated like qualitative data plus a statistical analysis, but we have to search for new ways to combine qualitative and quantitative research. One way to do this is to exploit the enormous amount of lexical items gathered through the inquiry. We start from the assumption that the morphology of the functional elements we observe must count for us to understand their properties as it must count for the child acquiring the language. The method we sketch out here is but one possibility and there certainly are others, but we try to provide an example of new ways of treating significantly big masses of linguistic data. When one looks at dialectal maps which also report the actual lexical items, and not only a symbol of the phenomenon, one cannot avoid noticing that some functional elements are extremely stable throughout the domain considered and other vary rather massively. For instance, in Poletto (2012) it is pointed out that the quantifier corresponding to 'all, everything' is etymologically very stable in the whole Italo-Romance and probably in the whole Romance domain, and still maintains the original Latin root (*totus*; cf. Rohlfs 1968: 228–229). One might think that the area of quantifiers is a rather stable one in terms of historical change of the lexemes, but if one looks at the forms for expressing the meaning of a quantifier like 'much', then the number of different lexical roots is overwhelming. Furthermore, it seems that the dialects have followed different conceptualization paths to encode the meaning 'much', some using a noun which can be associated to a big amount like 'mountain' or 'sack', while others take an adjectival form expressing simply an intensifier like 'thick' or even 'beautiful' or 'good'. We believe that this type of observations, when they are carried out on a big amount of data coming from historically very closely related languages, can be of help to understand the internal structure of functional elements and the type of semantic mechanisms that underlie these structures. Thus, it is possible to extract generalizations relevant for a syntactic analysis from raw lexical data simply by looking at the type and possible lexical variation for the same functional element starting from a simple but rather strong hypothesis: the *lexical variation index* of a functional word within a genetically related set of languages co-varies with the semantic and therefore syntactic complexity of the item itself. The more an element varies, the more it is complex; the less it varies, the simpler it is in semantic, and perforce in structural terms, since we take the internal layered structure of functional elements to reflect its semantic endowment. This means that a rather simple count of the possible lexical roots used in a set of related languages gives us very precise indications about the primitive components the functional element is made up of. To illustrate the point, we compare two different wh-items, i.e. two elements which in principle should be very similar, following Katz and Postal's (1968) basic intuition that all wh-items are made up by a Q-component followed by what we would nowadays call an ontological null category.[10] The first element is signaled by the wh- formant in English, and k- in Romance. The latter is a sort of classifier-like element providing some very general properties about the object in question, namely a silent item like PERSON, THING, PLACE, TIME, WAY, as recently proposed by Kayne in several works (among others 2005; 2006; 2012). So, according to this view, the wh-item *who* would correspond to 'which PERSON', the wh-item *what* to 'which THING', the wh *where* would be 'which PLACE' and so on. However, if we compare the wh-items corresponding to 'who' in the Italo-Romance domain, we obtain only four forms, three of which are explainable through phonological changes having to do with a general change in the vowel system (for instance lowering of the high vowel) or with the palatalization of the velar voiceless obstruent /k/. This means that, since phonological variants do not count when we establish the lexical variation index, essentially all

[10] We follow here Kayne's recent work on null nouns (see Kayne 2005).

varieties found belong to a single etymological type. In Table 1 we provide some examples from different dialectal areas.

Among the possible forms found through the whole ASIt database, some are clearly derived through phonological rules, like the palatalization from /ki/ to /tʃi/, or even the change into a fricative, similar to what has happened in French in non-functional words, so that we get /si/. We also find the lowering of the vowel from /ki/ to /ke/ or to /kə/ in those dialects that always require a final schwa even for functional elements. Other dialects have both phenomena, so in the variety of Arcola we find /tʃe/ with palatalization and lowering of the vowel. A real morphological difference is represented by forms like *kui* and *ku*, which probably derive from the oblique form of the wh-pronoun, not from the nominative, and *ki-ne*, which displays a reinforcing element (not in all contexts), typical of various functional elements like locative pronouns, pro-sentential negation, etc. A brief investigation shows that all these forms can be reduced to /ki/ through phonological rules and that the lexical variation index of the element *who* is actually only 3, i.e. the morphologically different forms are /ki/, /ku(i)/ and /ki/ plus the reinforcer *-ne*.

On the basis of the above, we would expect that a similar situation holds also for all the other wh-items, but if we consider the element meaning 'where', we find an astonishing difference. The same count we made for 'who' was also made on the basis of the AIS for the element meaning 'where' and this gives the astonishing result of 56 different forms only in the Northern Italian area, without counting the whole Central and Southern areas. Notice that 56 is the number we gather for forms which cannot be explained through phonological changes (see Munaro and Poletto 2014), but can only be derived through the presence of different formants/morphemes, as the pairs in Table 2 attest with respect to the formant *in* corresponding to the preposition *in*.

If one considers all the forms found in the AIS, the results can be represented as in Table 3, taken from Munaro and Poletto (2014). It is possible to identify at least 6 formants, represented in the first row.

Table 1: 'who' in Italo-Romance.

tʃi	Livigno (Sondrio, Northern Lombard); Ronzone (Trento)
kui	Cesarolo (Venezia, Western Friulian); Monasterace (Reggio Calabria); Rodoretto (Torino, Occitan)
kɔ	Liscia (Chieti, Abruzzese)
ku	Crotone (Calabrian); Barcis (Pordenone, Friulian); Monno (Brescia, Lombard); Biancavilla (Catania, Sicilian)
kine	Cosenza (Calabrian)
si	Erto (Pordenone, Friulian)
tʃe	Arcola (La Spezia, Ligurian); Triggiano (Bari, Apulian)
ke	Cairo Montenotte (Savona, Piedmontese/Ligurian)
kie	Bitti (Nuoro, Sardinian)

Table 2: *in-* in the form of the wh-item corresponding to 'where'.

Forms with *in-*	Forms without *in-*
Venice: indove	Teolo (Padua): dove
Parma: indo	Fiume (HR): do
Aldon (Verona): ando	Fiume (HR): do
Grado (Gorizia): indola	Ruda (Udine): dola
Poschiavo (CH): indond	Brescia: dund

Table 3: 'where' in the Northern Italo-Romance domain.

Formants						**Complete form**
in	***d***	***o***	***nd***	***là/v***	**Epenth. Vowel**	
am		w	(a)nd		a	amuanda
an	d	o		v	e	andove
(i)n	d	o		v	e	(i)ndove
in	d	o	nd			indond
	d	u	nd		a/i	dunda/dundi
n		u	nd		a	nunda
	d	o/u		v	e/a	dove/duve/duva
(i)n	d	o/u		v		(i)ndov
i	d	u		v		iduv
in	d	o		la		indola
		o	nd		e/a	onde/a
	d(a)		nd		e	dande
	d	u	n		(a)	dun(a)
in	d	u	n			indun
in	d	u	m			indum
n	t	u	n			ntun
		u	nt		a	unta
	d	u	nd			dund
n	d	u			a	ndua
(n)	d	o			a	(n)doa
n		u		w	a	nuwa
in	g	w			e	ingwe
in	g				e	inge
en	d	o				endo
an	d	u				andu
(i)n	d	o/u				(i)ndo/ndu
in	g	o/u				ingo/ingu
an	d	o				ando
an	d				e/a	ande/anda
en	t				e	ente
	d	o/u			e/a	doe/dua
	d	u/o		la		dula/dola
	d	u	n			dun
	d	u	m			dum
ne/a		w			a	newa/nawa
(i)n		o/u			a	(i)noa/(i)nua
an		w			a	anwa
n̩n	t					n̩nt
n	g					ng
an	t					and

(contd.)

Formants						Complete form
(a)n	t è (copula)					(a)ntè
in	d è (copula)					indè
		u/o		la		ula/ola
	d	o/u				do/du
	da		n			dan
		o	d			od
				v	e/i	ve/vi
n		o				no
agn		o				agno
	t		t		e/a	te/ta
a				lo		alo
	d		d		e/	de/da
		u				u
an						an
				la		La

The forms have been split according to the possible formants that have been identified in the varieties examined by Munaro and Poletto (2014). The table shows that the morphological complexity of the element meaning 'where' is much higher than the one of the element meaning 'who', thus contradicting the standard view. Furthermore, the formatives repeat themselves, and the 56 forms are actually the result of various combinations of five (or perhaps six, depending on how the final vowel is analyzed) formatives, some of which are clearly identifiable as either prepositions, (as *in*, or *d-*) or locative deictics (*lá*, corresponding to the distal 'there'). As for the comparison between the two wh-items, these tables indicate that the semantic and syntactic complexity of the two wh-words is rather different and that the element meaning 'where' has a much higher number of different formatives most probably because it has a high number of semantic features (and therefore syntactic projections). Munaro and Poletto (2014) further develop this observation as the basis to propose a layered structure of the wh-item corresponding to 'where' based on Cinque's (2010) proposal of the internal structure of locative PPs reported here.

(12) $[_{\text{PPdirsource}}$ 'from' $[_{\text{PPdirgoal}}$ 'to' $[_{\text{PPdirpath}}$ 'across' $[_{\text{StatP}}$ *a* $[_{\text{DegreeP}}$ 'two meters' $[_{\text{ModeDir}}$ 'straight' $[_{\text{AbsViewP}}$ 'south' $[_{\text{RelviewP}}$ *in* $[_{\text{DeicticP}}$ 'there' $[_{\text{AxPartP}}$ 'under/above/behind' $[_{\text{PP}}$ $[_{\text{P°}}$ *a*] $[_{\text{NPplace}}$ 'the table' $[_{\text{NP}}$ PLACE]…]

On the basis of the identification of the formatives presented in Table 3. Munaro and Poletto (2014) propose that the internal structure of the wh-item *where* is the following:

(13) $[_{\text{PPdirsource}}$ *da* $[_{\text{PPdirgoal}}$ *in* $[_{\text{PPdirpath}}$ $[_{\text{whP}}$ *o/u* $[_{\text{StatP}}$ $[_{\text{DegreeP}}$ $[_{\text{ModeDir}}$ $[_{\text{AbsViewP}}$ $[_{\text{RelviewP}}$ $[_{\text{DeicticP}}$ *là/v/nd* $[_{\text{AxPartP}}$ $[_{\text{PP}}$ $[_{\text{P°}}$ *a*] $[_{\text{NPplace}}$ e [PLACE]…]

Essentially, the idea is that the reason why *where* is lexically much more unstable than *who* is that it contains the whole complex structure of locative PPs, although only some of the internal projections can be lexicalized by formatives, as illustrated in (13). Furthermore, (13) does not lexicalizes only a subset of the projections in (12), but an intersection, since it contains a whP, which regular locative PPs do not contain. This means that in wh-items only some portions of the internal structure of locative PPs can

be actually lexically filled, but there is also something more. Without replicating Munaro and Poletto (2014) work, to which we refer for further discussion, the point here is that using etymological clues and the amount of lexical variation of a single functional item can give us precious insights into its internal syntactic structure, which in turn help us to identify some of its semantic components. A systematic comparison inside supposedly homogeneous classes of elements (like quantifiers, n-words, modals, different types of pronouns) might be one of the keys to understand more about the internal structure of functional items.

The lexical variation index can be thus seen as the first step of a method to identify formatives/morphemes that can make up a word. This type of procedure strongly recalls nanosyntactic work, but it actually does more, since it has the advantage of being able to identify the semantic value of the different formatives, an enterprise which is often not attempted in nanosyntactic work, where only the number and sequence of the morphemes is identified, but not their semantic value. This advantage precisely comes from the fact that we have at our disposal a set of closely related languages whose ancestor is a robustly documented and studied ancient language, so that their etymological value is rather easy to figure out and phonological changes are also well known, so that we are in a position to filter them out of the lexical variation index.

3 Concluding remarks

We conclude our methodological overview of different methodologies to treat big sets of data by noticing that big amount of data is often noisier than smaller set of data, where we can control for our experiment in a much more precise way, although all data contain a certain amount of noise, and it is easier to filter it when there are more. The general way to filter out this noise is to use statistical methods on the same type of evaluations we apply to smaller amounts of data. In this article we have tried to show that one can adopt a different perspective in treating big amounts of data so that instead of using the usual statistical evaluation methods, one can find a new combination of qualitative and quantitative research and extract different insights out of maps and tables simply looking at the distribution of phenomena, as in the case of the leopard spot pattern. Alternatively, it is possible to exploit big amounts of lexical variants to gather an insight into the syntactic layering and semantic components of functional items using purely distributional criteria. In other words, there are alternative, or better additional, strategies to the usage of refined statistical evaluations in trying to counterbalance the fact that all our data are noisy, in order to give a more precise account of dialectal and typological variation. We simply have to observe different types of distributional patterns as an indication of different classes of phenomena and use the evidence we have to formulate new types of questions.

Abbreviations

AUX = auxiliary, INF = infinitive, IMP = imperative, NEG = negation, PL = plural, PRT = particle, SG = singular, SCL = subject clitic

Acknowledgements

We thank Guglielmo Cinque, Franco Fanciullo, Cristina Guardiano, Romano Lazzeroni. Diego Pescarini, Silvia Rossi, Laura Vanelli and two anonymous reviewers for all the comments on the preliminary versions of this paper.

Competing Interests

The authors have no competing interests to declare.

Author Contribution

Jacopo Garzonio is responsible for sections 1, 2.1 and 2.2; Cecilia Poletto is responsible for sections 1.1, 2, 2.3 and 3.

References

Ascoli, Graziadio Isaia. 1873. Saggi ladini. *Archivio Glottologico Italiano* 1. 1–556.

Bartoli, Matteo. 1945. *Saggi di linguistica spaziale*. Torino: Rosenberg & Sellier.

Bayer, Josef & Ellen Brandner. 2008. On wh-head-movement and the doubly-filled-comp filter. In Charles B. Chang & Hannah J. Haynie (eds.), *Proceedings of the 26th West Coast Conference on Formal Linguistics*, 87–95. Somerville, MA: Cascadilla Proceedings Project.

Benincà, Paola & Cecilia Poletto. 2005. On some descriptive generalizations in Romance. In Guglielmo Cinque & Richard S. Kayne (eds.), *The Oxford handbook of comparative syntax*, 221–258. Oxford and New York: Oxford University Press.

Benincà, Paola & Cecilia Poletto. 2007. The ASIS enterprise: a view on the construction of a syntactic atlas for the Northern Italian dialects. *Nordlyd* 34. 35–52.

Berruto, Gaetano. 1987. Lingua, dialetto, diglossia, dilalia. In Günter Holtus & Johannes Kramer (eds.), *Romania et Slavia Adriatica*, 57–81. Hamburg: Buske.

Cardinaletti, Anna & Giuliana Giusti. 2016. The syntax of the Italian indefinite determiner *dei*. *Lingua* 181. 58–80. DOI: https://doi.org/10.1016/j.lingua.2016.05.001

Cerruti, Massimo & Riccardo Regis. 2014. Standardization patterns and dialect/standard convergence: A northwestern Italian perspective. *Language in Society* 43(1). 83–111. DOI: https://doi.org/10.1017/S0047404513000882

Cinque, Guglielmo. 1976. 'Mica'. *Annali della Facoltà di Lettere e Filosofia dell'Università di Padova* 1. 101–112.

Cinque, Guglielmo. 2010. Mapping Spatial PPs: An introduction. In Guglielmo Cinque & Luigi Rizzi (eds.), *Mapping spatial PPs: The cartography of syntactic structures* 6. 3–25. New York and Oxford: Oxford University Press.

Cornips, Leonie & Vincent de Rooij. 2014. Selfing and othering through categories of race, place, and language among minority youths in Rotterdam, The Netherlands'. In Peter Siemund, Ingrid Gogolin, Monika Edith Schulz & Julia Davydova (eds.), *Multilingualism and language diversity in urban areas: Acquisition, identities, space, education*, 129–164. Amsterdam and Philadelphia: Benjamins.

Dal Negro, Silvia. 2004. *The decay of a language: The case of a German dialect in the Italian Alps*. Frankfurt: Peter Lang.

Di Nunzio, Giorgio M., Jacopo Garzonio & Diego Pescarini. 2014. ASIt: Atlante Sintattico d'Italia. A linked open data geolinguistic web application. In Maristella Agosti & Francesca Tomasi (eds.), *Collaborative research practices and shared infrastructures for humanities computing*, 197–203. Padova: CLEUP.

Dorian, Nancy C. (ed.). 1989. *Investigating obsolescence studies in language contraction and death*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511620997

Garzonio, Jacopo. 2007. Complementatori nelle interrogative delle varietà trentine. In Gianna Marcato (ed.), *Dialetto, memoria e fantasia*, 179–183. Padova: Unipress.

Garzonio, Jacopo & Cecilia Poletto. 2017. Partitive objects in negative contexts in Northern Italian dialects. Ms. University of Padova.

Jaberg, Karl & Jakob Jud. 1928–1940. *Sprach- und Sachatlas Italiens und der Südschweiz*. Zofingen: Ringier.

Kayne, Richard S. 1992. Italian negative infinitival imperatives and clitic climbing. In Liliane Tasmowsky & Anne Zribi-Hertz (eds.), *Hommages a Nicolas Ruwet*, 300–312. Ghent: Communication and Cognition.

Kayne, Richard S. 2005. *Movement and silence*. Oxford/New York: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780195179163.001.0001

Kayne, Richard S. 2006. On parameters and on principles of pronunciation. In Hans Broekhuis, Norbert Corver, Riny Huybregts, Ursula Kleinhenz & Jan Koster (eds.), *Organizing grammar: Linguistic studies in honor of Henk van Riemsdijk*, 289–299. Berlin and New York: Mouton de Gruyter.

Kayne, Richard S. 2012. A note on grand and its silent entourage. *Studies in Chinese Linguistics* 33(2). 71–85.

Ludlow, Peter. 2011. *The philosophy of generative linguistics*. Oxford and New York: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199258536.001.0001

Luraghi, Silvia. 2012. Partitives and differential marking of core arguments: A crosslinguistic survey. Unpublished manuscript, University of Pavia.

Lusini, Sara. 2013. *Yes/No question-marking in Italian dialects. A typological, theoretical and experimental approach*. Leiden: University of Leiden dissertation.

Manzini, M. Rita & Leonardo M. Savoia. 2011. *Grammatical categories: Variation in Romance languages*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511974489

Munaro, Nicola & Cecilia Poletto. 2014. Synchronic and diachronic clues on the internal structure of 'where' in Italo-Romance'. In Paola Benincà, Adam Ledgeway & Nigel Vincent (eds.), *Diachrony and dialects*, 279–300. Oxford and New York: Oxford University Press.

Obenauer, Hans-Georg. 2004. Nonstandard wh-questions and alternative checkers in Pagotto. In Horst Lohnstein & Susanne Trissler (eds.), *Syntax and semantics of the left periphery (Interface Explorations 9)*, 343–384. Berlin and New York: Mouton de Gruyter.

Oliviéri, Michèle & Cecilia Poletto. 2018. Negation patterns across dialects. In Mirko Grimaldi, Rosangela Lai, Ludovico Franco & Benedetta Baldi (eds.), *Structuring variation in Romance linguistics and beyond. In honour of Leonardo M. Savoia*, 131–146. Amsterdam and Philadelphia: Benjamins.

Parry, Mair. 2003. Cosa ch'a l'é sta storia? The interaction of pragmatics and syntax in the development of WH-interrogatives with overt complementizer in Piedmontese. In Christina Tortora (ed.), *The syntax of Italian dialects*, 152–174. Oxford and New York: Oxford University Press.

Poletto, Cecilia. 2012. Contrastive linguistics and micro-variation: The role of dialectology. In Matthias Hüning & Barbara Schlücker (eds.), *Contrastive linguistics and other approaches to language comparison*, 47–68. Amsterdam and Philadelphia: Benjamins.

Poletto, Cecilia & Laura Vanelli. 1995. Gli introduttori delle frasi interrogative nei dialetti italiani settentrionali. In Emanuele Banfi, Giovanni Bonfadini, Patrizia Cordin & Maria Iliescu (eds.), *Italia settentrionale: Crocevia di idiomi romanzi. Atti del convegno internazionale di studi. Trento, 21–23 ottobre 1993*, 145–158. Berlin and New York: De Gruyter.

Rohlfs, Gerhard. 1968. *Grammatica storica della lingua italiana e dei suoi dialetti*: *Morfologia*. Torino: Einaudi.

Stark, Elisabeth. 2008. The role of the plural system in Romance. In Ulrich Detges & Richard Waltereit (eds.), *The paradox of grammatical change: Perspectives from Romance*, 57–84. Amsterdam and Philadelphia: Benjamins.

Vai, Massimo. 1998. Imperativo, negazione e clisi tra latino e neolatino. In Paolo Ramat & Elisa Roma (eds.), *Sintassi storica: Atti del XXX congresso internazionale della SLI*, 647–670. Roma: Bulzoni.

Zanuttini, Raffaella. 1997. *Negation and clausal structure: A comparative study of Romance languages*. Oxford and New York: Oxford University Press.

How to cite this article: Garzonio, Jacopo and Cecilia Poletto. 2018. Exploiting microvariation: How to make the best of your incomplete data. *Glossa: a journal of general linguistics* 3(1): 112. 1–21, DOI: https://doi.org/10.5334/gjgl.556

Submitted: 27 October 2017 **Accepted:** 22 July 2018 **Published:** 22 October 2018

]u[*Glossa: a journal of general linguistics* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS