Dialectologia 25 *(2020), 59-86.*
ISSN: 2013-2247

Received 21 January 2020.

Accepted 27 March 2020.

UNIFYING ANALYSES OF MULTIPLE RESPONSES[1]

Gotzon AURREKOETXEA[1], John NERBONNE[2] & Jesus RUBIO[3]

UPV/EHU[1,3] * / Universities of Groningen and Freiburg[2] *

gotzonaurre@gmail.com / j.nerbonne@rug.nl / jesusangel.rubio@ehu.eus

Abstract

In dialectology we often encounter irreducible *variation* in its data, i.e., multiple responses to its probes about the form of a word or phrase. Dialectometry seeks to measure the differences between dialects and has developed several ways to measure the difference between responses when one or both of them is non-unique. We introduce here BILBAO DISTANCE, where the cardinality of response is unimportant, which may be combined with various weighting functions such as edit distance or inverse frequency weighting, and which yields intuitively appealing measures, e.g., when applied to a singleton set {a} and a set with the same element plus a second, yields d({a},{a,b}) = 0.5. It overcomes flaws in earlier proposals and is conceptually simpler and computationally more efficient to apply than earlier measures. We suspect that its results satisfy the metric axioms, as it is certainly symmetric and measures the difference between identical sets as zero.

Keywords

dialectology, dialectometry, local variation, multiple responses, multiple values

[1] Aurrekoetxea suggested the paper and wrote §1.2.2. Nerbonne wrote §1.2.1 and section 2. They collaborated on the other introductory sections, background and remarks on other atlases. The novel measure of the difference between cells with multiple values is due to Rubio, who also collaborated with Nerbonne in writing section 3.

* Faculty of Economics and Business (Sarriko). Avenida Lehendakari Agirre, 83 - 48015 Bilbao, Spain.

* Germanic Linguistics, Belfortstr. 14, Albert-Ludwigs-Universität, D-79085 Freiburg im Breisgau. Germany.

UNIFICANDO EL ANÁLISIS DE RESPUESTAS MÚLTIPLES

Resumen

En dialectología a menudo encontramos variaciones irreducibles en sus datos, es decir, múltiples respuestas o múltiples formas de una palabra o frase en una misma localidad. La dialectometría como disciplina que mide las diferencias entre los dialectos ha desarrollado varias formas de medir la diferencia entre las respuestas de una o más localidades no son únicas. Presentamos aquí la DISTANCIA BILBAO, en la que la cardinalidad de la respuesta es intranscendente, que se puede combinar con varias funciones de ponderación, como la distancia de edición o la ponderación de frecuencia inversa, y que produce medidas intuitivamente atractivas, por ejemplo, cuando se aplica a un conjunto único {a} y un conjunto con el mismo elemento más un segundo diferente, produce d ({a}, {a, b}) = 0.5. Esta nueva unidad de distancia supera defectos de propuestas anteriores, es conceptualmente más simple y computacionalmente más eficiente de aplicar que aquellas. Creemos que sus resultados satisfacen los axiomas métricos, ya que ciertamente es simétrico y mide la diferencia entre conjuntos idénticos como cero.

Palabras clave

dialectología, dialectometría, variación local, respuestas múltiples, múltiples valores

1. Background

Dialectology has benefited from a long tradition of systematic data collection, and from the laudable convention of recording the results of that collection in large data atlases, often accompanied by even more extensive databases on which the atlases are based. This has inspired a tradition of analyzing such collections or databases quantitatively, which by Séguy (1973) initiated, and which has become known as dialectometry (see Wieling & Nerbonne 2015 for a recent survey of work in this direction).

Geographic variation is sometimes referred to as DIATOPIC, and social as DIASTRATIC; they may be mentioned together with DIAPHASIC variation, the variation due to style and context and even diachronic variation, focused on the variants (changing) with respect to time. Especially since Labov (1969) proposed that variability is inherent in language, the studies of variation have progressed from being predominantly diatopic to studying

diastratic and diaphasic variation as well. Although there is no standard view of language variation in formal theoretical models (Hinskens 2018: 89), it has gradually come to be assumed that variation is intrinsic to the system. And if variation is intrinsic, it is natural that it be also part of individual performance. We may assume therefore that there is intra-speaker as well as inter-speaker variation (Honeybone 2011: 156-176), just as variationist linguistics has long argued. And where there is variation there are co-occurrences of different forms expressing the same linguistic content, which means that different responses may be made to field-workers' questions, i.e. MULTIPLE RESPONSES (MRs), leading to multiple values in data collections, something we also refer to as POLYMORPHISM.

The assumption of this variation is not new in dialectological works. In fact, this type of variation was referred to as polymorphism in traditional works on linguistic variation. Allières (1992: 187), one of those who has investigated the subject the most, flatly affirms that polymorphism belongs to the system and not to the competence of the speakers. The same author regrets that there has been negligence in the treatment of this type of variation, using a quote by K. Jaberg taken from his *Der Sprachatlas (l'AIS) als Forschungsinstrument:*

> Daβ dasselbe Wort je nach den Umständen sehr verschieden ausgesprochen werden kann, dürfte nach den systematischen Untersuchungen von Rousselot, Gauchat, Terracini, Bloch, Lutta und nach den Beobachtungen von vielen anderen Gelehrten auch außerhalb des romanistischen Gebiets theoretisch kaum mehr geleugnet werden, trotzdem auch heute noch in praxi viele Dialektforscher konsequent an diese Tatsache vorbeigehen, die für eine saubere Einordnung der Beispiele in die Paragraphen einer historischen Lautlehre sehr unbequem ist (Allières 1992: 181).

There have been different proposals for how to analyze (and quantify) the differences between such sets of multiple responses. We review these here, criticizing some, and proposing a novel, but not too radical alternative.

We are aware that data science has suggested a number of ways to quantify the differences (and/or similarities) between structured elements based on their components

(Manning & Schütze 1999), and we review a number of these in the course of the paper (see Sec. 2.1 and 2.3).

1.1 Multiple responses, multiple values

In collections of categorical and other non-numerical (see below) data, it is common to find that some data fields normally recording a single value are in fact occupied by two or more, apparently indicating that each of the alternative values would be acceptable. For example, if we record information about subjects in a typical university experiment with student participants, we might record the major field of study as 'Linguistics', 'Mathematics', etc., but there will be students who provide two and perhaps more responses, e.g. 'Linguistics' and 'Psychology'.

This also happens frequently in collections of linguistic data, our focus here. We are often interested in characterizing the differences between two or more varieties, and this is done by comparing the differences in a large paired sample. We then compare item by item in the sampled set. If items have single values, then it is straightforward to note whether items are the same or different and then to examine the total number of different items or identical ones, and this has been the standard procedure in quantitative dialectology since its inception (Séguy 1973; Goebl 1984).

We wish to emphasize that the frequencies of the elements in the items we are interested in comparing are too small to be reliable indications of population frequencies. If we were dealing with larger frequencies, say twenty or more per item, we might resort to statistical techniques developed for the comparison of frequencies such as chi-square, odds ratios or logistic regression (Paolillo 2018). But when we see two responses *cup* from one group of respondents and three from another, then the frequency comparisons cannot be reliable.

1.2 MRs in Variationist Linguistics

VARIATIONIST LINGUISTICS is the name given to the joint venture of dialectology and sociolinguistics (Chambers & Trudgill 1998) that studies how languages vary. The former,

and older discipline studies how languages vary with respect to geography, and the latter how they vary with respect to social standing. As noted above, we have, not only geographic (or diatopic) variation but also diastratic (social) and diaphasic variation (i.e., that due to style and context) and even diachronic variation, focused on the variants (changing) with respect to time. We begin somewhat heavy-handedly, belaboring the point that linguistic variation is multi-faceted, in order to hammer home the point that surveys of dialect variation are quite likely to encounter local variation exemplifying any of these factors, perhaps in combination. Whenever there is genuine local variation, language surveys will record more than one response for a given question or probe. In modern London English one might record words with a final /t/ as a glottal stop [ʔ]: *not* as [nɔʔ], rather than [nʔ t^{-}] or [nʔ t^{h}]. The point that local variation is endemic should need no elaborate argument when considering pronunciation in language surveys (but see below for discussion of some cases).

We wish to present in this paper an improved way to measure the difference between survey items involving multiple responses, so we likewise wish to insist that the phenomenon is genuine. Our respected colleague, Prof. Hans Goebl, has argued that multiple responses are survey artefacts reflecting poor data collection techniques (1997: 28). In particular he has speculated that multiple responses may reflect social (diastratic) differences that ought to be controlled for more strictly during dialectological field work.

We document below (in Sec. 1.2.1-1.2.3) how existing data sources often include multiple responses, which means that we wish to have a means of analyzing them. If we discarded all such sources, variationist linguistics and dialectology in particular would lose treasure troves of data, and this would be unacceptable.

Variation exists not only within entire predominantly monolingual countries, but also within settlements, and even within the speech of individuals (as noted above). Variation within individuals may reflect diaphasic influences of style or context, but it often reflects the fact that individuals are masters of more than a single variety. This possibility implies that there must be variation in the speech of individuals. Many speakers are proficient in modern standard languages, but they adjust easily to regionally colored varieties of these (regiolects), and they adjust in both varieties to formal and

informal situations (diaphasic abilities). It is only natural that field workers encounter multiple responses especially in the speech of such individuals.

Nor do we wish to suggest that variation is limited to pronunciation. In lexical surveys the existence of variation means that one often cannot say which word is *the* form used at a particular location. A survey might record either *begin* or *start* for term indicating the initial phase of an action, or perhaps *difficult* or *hard* for the opposite of *easy*. We document more examples below. Morphological variation is common enough to make its way into handbooks and standard grammars. Most grammars list *shaven* and *shaved* as alternative past participles of *shave*, and likewise *shorn* and *sheared* as alternative participles of *shear* (even if they indicate a preference for *shorn* and *shaven* used adjectivally). Google lists 200K hits for "closely shaved" and 100K for "closely shaven". Syntactic variation is likewise commonplace.

It has been recognized widely in linguistics (Labov 1969) that all varieties admit variation even while respecting enough invariance in language to guarantee communication Variation is an essential element of languages, because it provides a means for speakers to signal their identification with some regions, groups and even styles within a language area.

We would nevertheless like to note examples of language surveys where multiple responses are frequent, and where ignoring them would expose analyses to criticism. While we suspect that it is very common to prefer single responses, which are, of course, easier to analyze, multiple responses are frequent and probably often reflect linguistic behavior accurately.

1.2.1 Multiple responses in LAMSAS

The *Linguistic Atlas of the Middle and South Atlantic States* (LAMSAS) includes material collected from 1933 through 1974 on the eastern seaboard of the US. More than 70% was collected by a single field worker, Guy Lowman, who unfortunately died in an automobile accident during his work. Raven McDavid completed most of the rest of the interviews much later. The collection protocol specified that the lexical realization of 151 concepts be recorded, using questions such as "*If the sun comes out after a rain, you say*

the weather is doing what?", to which responses such as *clearing up*, *fairing off* and forty other dialect variants were recorded (Kretzschmar, McDavid, Lerud & Johnson 1994). As earlier work has documented, multiple responses were very common in this data collection effort (Nerbonne & Kleiweg 2003). Table 1 indicates that multiple responses were common in all of the field workers' interviews, but especially in McDavid's, where there were 1.3 responses per concept on average. In fact, students of dialectology also appreciated the large number of responses that McDavid often obtained, reasoning that he had collected more of the variation that was present (W. Kretzschmar, personal communication).

Fieldworker	Number of Interviews	Number of Responses	Mean Responses/ Interview	SD Responses/ Interview
Lowman	826	123,990	150.1	25.3
McDavid	278	54,855	197.3	76.8
others	58	12,057	207.9	43.9
Totals	1162	190,902	164.3	49.6

Table 1. LAMSAS response rates. Field worker Guy Lowman conducted over 70% of the LAMSAS interviews, and Raven McDavid 24%, leaving only 5% for others. Note that while Lowman obtained an average of about one response per question, McDavid and others obtained an average of over 1.3 responses per concept, and the standard deviations in the number of responses was much higher for McDavid and the other field workers than it was for Lowman. We take these figures to demonstrate that multiple responses were common in this linguistic survey. Note further that, given the standard deviation in Lowman's response numbers, his interviews often contained multiple responses as well. The table is repeated from Nerbonne & Kleiweg (2003).

1.2.2 Multiple responses in Basque Linguistic Atlas-EHHA

Just as in many other language atlases, the Basque linguistic atlas (hence EHHA, from *Euskararen Herri Hizkeren Atlasa)* has often collected data involving multiple responses. Multiple responses are given in all linguistic categories, both in the lexicon and in grammar. In this linguistic atlas the directors gave precise instructions to the collectors

to investigate each concept and collect all possible answers. Below we show a sample of 51 questions on the nominal declension of Basque in which MRs were recorded for every query (Figure 2).

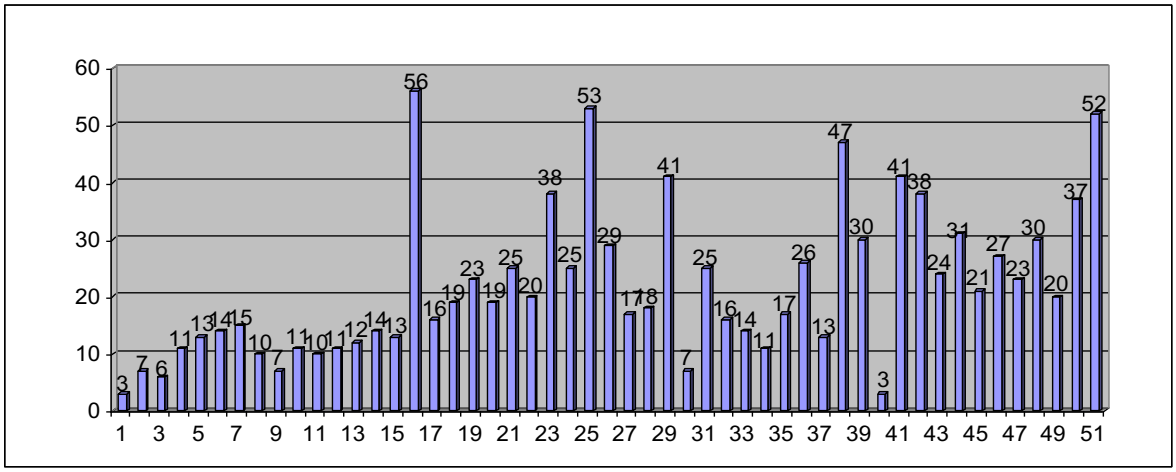

Figure 2: Number of localities with MRs in 51 questions on declension in EHHA (Videgain & Aurrekoetxea 2015: 123)

The figure shows the questions (from 1 to 51) and the number of localities with MR in each one. As one can see, there is no question to which every locality offered only one answer. The question with the most locations with MR is question 16, in which multiple responses were found in 56 locations. Some cases of MR are discussed in the following paragraphs.

A great variety of MRs has been found in the case of Basque. To give a sample of them we will begin with responses of different etymologies, constructed with different morphemes.

Words with different etymology have been considered (dialectal) synonyms in the scientific literature. An example of what we are saying are the words *urde* and *zerri* 'pig' (EHHA IV: map 779), collected in Makea. Although it is known that in localities where both words are used they may sometimes have different semantic nuances, in Makea they are synonymous. Two words that are native for local speakers and have different origins are used interchangeably for the same meaning. The same happens in Arboti and other locations for the words *urde* and *borthako*. Therefore, since two words with different

etymologies are used in the same locality for the same meaning and are regarded as are native, they are multiple responses.

Regarding the origin of the word, a loan is sometimes used next to an indigenous word. Regardless of the frequency of these words, if in the same locality the same or different speakers are found to use two words, we are faced with an MR. For example, in Zeanuri the words *mariposa* and *eskabi* (EHHA I, map 25) have been elicited for the concept 'butterfly'; the first a loan and the second a word of native origin. This concept has given rise to a plethora of responses in many localities throughout the Basque territory. In some of them up to four responses have been collected, as in Leitza *(mariposa, tximeleta, pinpilinpausa* and *mitxirrika);* the first is a loan from Spanish, and the remaining are indigenous.

Words composed of different morphemes have also been collected, such as *apo* and *txerrapo* 'boar' (EHHA IV: map 781); the first with a single morpheme and the second with two (txerri + apo 'pig + boar'), have been collected in the western and central part of the Basque regions for the concept 'boar, male pig'. Another example of what we are discussing is presented by the pair *txarrikorta* and *txarritegi* (EHHA IV: map 786) for the concept 'pig sty', collected in Dima, in which, although the first morpheme is identical in both words, the second morphemes “- korta” and “-tegi ” have the same semantic value, 'stable'. In Orozko three words have been collected *txarrikorta, txarritegi* and *txarrikortatxi* (first element “txarri-” and as second “-korta”, “-tegi”, already seen, and “-kortatxi” that has the same meaning). Finally, in certain eastern towns (Mugerre, Uztaritze, Garrüze, etc.), *kosta* and *kostaleta* (EHHA IV: map 799) have been collected for 'spare ribs', which, as one can see, have the same first “kosta-” morpheme while differing in the second "-leta". Despite the plentiful evidence of morphological polymorphism, MRs based on phonetic divergences are more abundant. On map 779 mentioned above, the variants *zerri* and *txerri* have been registered in central locations such as Getaria, Hondarribia, Lasarte-Oria, Eugi and Ezkurra, apparently with the same meaning, although in other locations they are registered with different meanings.

In some localities specific words have been lost and replaced with generic words as if they were synonyms. This is the case, for example, *bero* and *irausi* for 'heat of pig'

collected in Mendaro (EHHA IV: map 782): the first is a generic word that means 'heat' (sexual receptiveness) and is also used for other animals, the second specific i.e., only used for pigs. It is a phenomenon that usually occurs in specific names of traditional trades that are gradually lost.

It is much more difficult, and at the same time more doubtful, to ascertain MRs given an imprecise semantic concept, such as those in the colloquial language *bihurritu, atera, urten*... for the term 'dislocate'. Although the meaning of these words in dictionaries is very precise, in the colloquial language and during the work of linguistic surveys it is more difficult to determine if the informant has used the usual term in their locality or has used a term in an *ad hoc* way.

Grammatical elements are not immune to polymorphism in oral language. Not only in nominal and verbal morphology, but also in syntax and in phonology, a number of different forms with the same meaning have been collected.

In nominal morphology, and more specifically in declension, MRs appear more commonly than is usually appreciated. The EHHA reader will find abundant evidence for this in volume VI. As examples we cite only a few: In Zeberio, but also elsewhere, plural possessive genitive forms *sémean* and *semén* have been observed, in Busturia *semen* and *sémien*, and in Bardoze *semeén* and *semén* 'of the sons' (EHHA VI: map 1046). In all cases, at a glance, they are examples of a progressive loss of the indeterminacy feature.

The second example concerns the case '-arengan' (animate, inessive case), a feature in decline in colloquial language. In Mañaria, *alábias* 'with the daughter' (sociative case, also known as concomitative case) and *alábiagán* 'at the daughter' (inessive) and other locations similar responses have been collected in the western part of the Basque country. In many other locations, responses have been collected only with the sociative case (Arrieta, Busturia ...). Without going into an analysis of the factors that favor the use of the sociative for animate inessive, here we will only highlight the fact that the same phrase has been realized using different cases, which apparently reflect no meaning difference for the speakers.

Something similar happens with some verbs: there are cases of phonetic variants and there are cases involving the use of different verbal forms, or even the use of conjugated and unconjugated forms that have been observed in response to the same

question. Such is the case of Ataun in which the verbal forms *nun* and *non* (EHHA VI, map 1339) have been collected in the third person singular past tense of the auxiliary used with transitive verbs of two arguments asking to translate the sentence 'I had a house', in Dima *neuen* and *nauen*, in Zeberio *nendun* and *nindun*, in Hondarribia *nuen* and *nun*, etc. In all cases they are phonetic variants of the same verb. There are also cases of paradigm change, as in Sondika with *nun* and *nendun*, in Lemoa *ninduen* and *ostén* (EHHA VI, map 1346), in the latter case with a verb change from two arguments to three. The use of unconjugated forms is a common resource in the colloquial oral language to replace subjunctive forms; in Amezketa, *arrimatzea auke* (unconjugated) and *dakiola* (EHHA VI, map 1325) have been recorded for the auxiliary for intransitive verbs of two arguments (verbs with no accusative): the first is a form that is gaining popularity in the central area.

Nor is syntax exempt from MR. Two cases illustrate this: in Hendaye, *gaten* and *gaterat* (nominalized form in two declension cases) have been observed as alternative verbal complements to the verb “utzi” 'leave (that ...) / leave doing ...', when the informant was request to translate ‘let him/her do that’ (EHHA VII, map 1749); in Etxarri-Larraun, *eztot eose obik* [neg + verb + object] and *obik eztot eose* [object + neg + verb] have been observed when inquiring about the structure of a negative sentence (EHHA VII, map 1752).

As expected, the MR in phonetics are innumerable. In many other cases in Landibarre, for example, *partitzeat* and *egiterat* have been collected, in Irisarri *partitzerat* and *partitzeat* (with an intervocalic -r- drop) (Aurrekoetxea 2018).

1.2.3 Other language surveys

Blanquaert & Peé edited the *Reeks Nederlandse Diallectatlassen* (RND), the (older) standard collection of Dutch dialect surveys, in which 141 sentences are recorded (Blanquaert & Peé 1925-82). Wilbert Heeringa digitized 125 words from these, justifying his choice of words and treatment of factors such as sandhi, syllable reduction, and the use of diminutive vs. stems (Heeringa 2001). He notes that varying lexical choices were included, sometimes with an indication that one lexicalization was archaic, but often

without any such specification, in which case he included both in his studies of lexical and pronunciational variation (Heeringa 2001: Sec. 6.2).

Taeldeman & Goeman (1996) report on a much larger (and newer), more systematically collected Dutch data set (the Goeman-Taeldeman-van Reenen set, GTRP), which also includes multiple responses, as Wieling, Heeringa & Nerbonne (2007) note, although they present no statistics about how often this occurs. In a large selection used to demonstrate the Gabmap web application, no alternate pronunciations are recorded (www.gabmap.nl/~app/examples/).

The *Atlas Linguistique du Gabon* (ALGAB) is an atlas collected at Lyon's *Laboratoire dynamique du Language* (Hombert 1990). The data contains numerous multiple responses and has been analyzed dialectometrically (Alewijnse, Nerbonne, Van der Veen & Manni, 2007; Manni & Nerbonne, to appear) and is available as a Gabmap demonstration set (www.gabmap.nl/~app/examples/).

Another linguistic atlas that collects multiple responses is the *Sprachatlas von Bayerisch-Schwaben* (SBS) compiled under the direction of Werner König at the University of Augsburg. An interesting dialectometric analysis of this atlas has been carried out by Rumpf, Pickl, Elspaß, König & Schmidt (2009).

There are examples of multiple answers in *Atlas linguistique et ethnographique du Languedoc Occidental-ALLOc,* compiled by Ravier (1978-1993). Ravier developed a survey methodology in which, apart from the speaker's responses to the questions asked by the interviewer, he collected the words accepted by the speaker that had been proposed to him by the interviewer. In this way he inquired about the passive lexicon of the locals.

In *Atlas Lingüistico Galego,* ALGa, multiple responses have also been collected. As Sousa (2017: 12) puts it, "As often occurs in geolinguistic projects based on the use of questionnaires, the ALGa researchers collected at certain times more than one answer to questions from the notebook. Multiple answers are quite frequent in the section dedicated to lexicon."

Another good example is the *Atlas Lingüístic del Domini Catalan*-ALDC atlas, led by Veny & Pons i Griera (2001-2018). In practically all the maps the existence of multiple responses can be confirmed. In de same linguistic domain, Perea (2009) accounts for the data collected by Alcover for *La flexió verbal en els dialectes catalans* (1929-1933), in

which she records abundant multiple responses. This compilation of the verbal forms of Catalan is a good demonstration of the changing state of the oral language.

2. Toward a treatment

There are many possible treatments, perhaps the simplest of which is the following. When comparing two items potentially containing multiple values, simply pick one of them at random and use that as a representative for the item in question. So if the first cell contains {a, b, c} and the second {c, d} and one picks a and d respectively, then the comparison will result in no overlap. If on the other hand one picks c from both sets, then there is an overlap, and the items will be counted as the same. If many items are included in the sets being compared, then the inaccuracy of choosing at random should be tolerable, but we still dislike this procedure, first because we think that the multiple responses reflect linguistic reality, and we prefer not to simply ignore the fact that multiple responses are given, even if this is only done stochastically. Second, we are often interested in the contribution of an item to an aggregate analysis, for example, when we ask which items are most typical of a dialect area we have determined in the aggregate analysis (Prokić, Çöltekin & Nerbonne 2012). Third, we may be interested in what concepts tend to be lexicalized in various ways, a question Franco (2017) addresses from the perspective of cognitive linguistics. Indeed, lexical diversity, and which concepts tend to display it, is the focus of Franco's work. From the second and third perspective, case, the stochastic procedure threatens to hide what is distinctive about the item. Still, this "quick and dirty" approach may be sufficient for some purposes.

Before reviewing proposals, it will be worthwhile distilling some conditions which a good procedure ought to fulfill. First, since we cannot rely on the how often a response occurs in the data (see discussion above), this should not matter in the solution at all. Mathematically, we may encounter a multiset in our data, but we are not interested in how often individual elements occur. So the difference between {a, a, b, c, c, c} and {a, b, d, d} should be the same as the difference between {a, b, c} and {a, b, d}, i.e. d({a, a, b, c,

c, c}, {a, b, d, d}) = d({a, b, c}, {a, b, d}). We can safely apply the procedure to the set any multiset reduces to.

Second, the distance between identical sets ought to be zero, just as for any distance metric. In particular d({a, b}, {a, b}) = 0.

Third, it is appealing to wish to see the distance between two sets {a,b} and {a,c} to be equal or less than 0.5. Note that the two sets have one of their two elements in common, which is to say that they should have, as a minimum, a half in common, or, put differently, they should have, as a maximum, a distance of 0.5. On the other hand, the distance between two sets {a,b} and {a,c} should arguably be strictly greater than the distance between sets {a,b} and {a}, where c element is different to either a and b. As a consequence, the distance between a two-element set and either of its one-element subsets should be strictly less than 0.5 (i.e. $d(\{a,b\}, \{a\}) < d(\{a,b\}, \{a,c\}) \leq 0.5$).

Fourth, we prefer procedures that can generalize to situations in which the difference between data items is mediated by a cost function, i.e. a weighting. One such weighting is a string edit distance, used to gauge the pronunciation difference between words as they are pronounced at different dialect sample sites (Nerbonne 2003). Since we have published a good deal on this and its validation, we will not present the edit distance measure here, but instead refer the reader to earlier work (Nerbonne & Heeringa 2010; Heringa, Nerbonne & Kleiweg 2002).

Another such cost function is common in quantitative dialectology, namely the inverse frequency weighting Prof. Goebl (1984) introduces as 'weighted identity' (*gewichteter Identitätswert*). The idea is appealing as can be appreciated in an example. If I ask the LAMSAS question "*If the sun comes out after a rain, you say the weather is doing what?*", and obtain 400 instances of *clear* or *clearing up*, but only 20 instances of *fairing off,* then the one sort of sameness is more impressive than the other. Then the fact that two respondents use *fair off* may be regarded as a stronger indication of affinity (than the fact the two use *clear* or *clear* up). We emphasize that the frequency used in the weighting is that within an entire data set, overall frequency, not just within the items being compared. Goebl (2010) justifies the weighting noting that "rare and therefore 'more important' language features should be privileged over frequent ones, as they might be considered 'trivial'".

2.1 Jaccard, Manhattan distances

JACCARD DISTANCES (Manning & Schütze 1999: 299) are often used to gauge the distance between sets, and we are comparing sets of responses. The definitions are as follows:

$$\text{Jacc-sim}(A,B) = |A \cap B|/(|A \cup B|)$$
$$\text{Jacc-diff}(A,B) = 1 - |A \cap B|/(|A \cup B|)$$

But we wish to apply the cost function (see last section) to the ordered pairs of elements it is defined on, but both the numerator, $|A \cap B|$, and the denominator, $|A \cup B|$, quantify sets of singletons. They therefore do not seem to offer the flexibility we are looking for. Of course, this might be a fine approach to try if no cost function is involved.

MANHATTAN DISTANCE (also known as the L_1 distance) is defined as *Manh-Dist(A,B)* = $\sum_i |A_i - B_i|$, i.e., for each dimension *i*, we sum the differences in cardinality between A and B (Manning & Schütze 1999: 304). In our case the dimensional index *i* would be used to range over the different (categorical) responses. If we applied this to the example used above, the *Manh-Dist({a, b, c} {a, b, d})* would be 2, since there is an a and a b in each set, which therefore contribute zero to the distance. However, c and d each contribute one in distance since they are present in only one of the sets. It is straightforward to apply this idea to obtain a difference measure for multi-sets, i.e. sets where elements are associated with cardinalities, and Nerbonne (2017) discusses its potential as a means of "quantifying differences in histograms". We argued in section 2 (above) that we should ignore the frequencies of responses at a given site, since the frequencies are normally too low to be reliable. But a generalization of Manhattan distance suggests itself, in which one replaces cardinality with a cost function such as inverted overall frequency or edit distance. We return to this in Section 3 below.

2.2 Covering sets

Nerbonne & Kleiweg (2003) faced the problem of evaluating the distance between multi-valued items in an analysis of the LAMSAS lexical data.[2] They were aware that a simple definition based on the cross-product (also known as Cartesian product) of the multiple responses would run into trouble. If we examine the cross product of items A and B, i.e. A X B, and further use the mean of the differences in A X B, then we risk not evaluating the distance between {a, b} and {b, a} as zero, since, after all, A X B = { <a,a>, <a,b>, <b,a> and <b,b>} , and the d(a,b) = d(b,a) = 1. This simplistic "lifting" of the difference function to the mean of the differences of the cross-product would thus gauge d({a, b}, {b, a}) not to be zero, but rather as 0.5. In addition, we wished to ignore multiple occurrences of the same value, e.g. in cells such as $\{a, a, b, b, b\}$, which will be evaluated as $\{a, b\}$.

2.2.1. Minimal cost covers

We began by reducing the multiple values we might find in records to true sets, i.e. with no repeated elements. The original collection might be a multiset reflecting the results of a survey. As argued above, however, we are focused on the case where the cardinality of the elements in a given group of responses is not to be evaluated. Normally there are not enough responses for this to be done reliably, so we ignore multiplicities greater than one.

Nerbonne & Kleiweg (2003) then borrow the notion of a PROJECTION of a set of *n*-tuples at the *i*-th position from database theory. The *i*-th projection of a set of of *n*-tuples is just the set of all elements that occur in the *i*-th position in the set of *n*-tuples. Given a set *S* of *n*-tuples, $proj_i(S) = \{s_i \mid < s_1 \ldots s_i \ldots s_n > \in S\}$, where $1 \leq i \leq n$. So given S={<a,a>, <a,b>, <a,c>,<b,c>}, then $proj_1$(S)={a,b} and $proj_2$(S) = {a,b,c}. Of course, we are interested in the pairs formed by comparing A and B.

[2] We were once asked whether the introduction to the covering-set treatment, which described looking for minimally distant items first in the one set, then in the other, and then taking a mean of the sum of these minimal distances, was intended as a second possible solution. But it was intended as an informal procedural description of the covering-set approach.

They then introduce the notion of a COVERING SET of ordered pairs. Given a set *S* of ordered pairs, a subset S' ⊆ S covers S if its first projection is the first projection of S and likewise its second projection is S's second projection.

$$S' covers\ S \ \stackrel{\text{def}}{=}\ S' \subseteq S \wedge Proj_1(S') = Proj_1(S) \wedge\ Proj_2(S') = Proj_2(S)$$

Note that covering sets are always subsets of the original set of pairs, so that they only contain pairs resulting from matching the two sets of values (into the cross-product). But note, too, that they may be smaller than the full cross-product. So if A={a,b} and B={a,b,c}, then A X B ={<a,a>, <a,b>, <a,c>,<b,a>, <b,b>, <b,c>}, with a cardinality of six. But, as we noted above, there also exists S'={<a,a>, <a,b>, <a,c>,<b,c>}, where *proj*$_1$(S')={a,b} and *proj*$_2$(S) = {a,b,c}, which, of course are the first and second projections of A X B. So S' also covers A X B. It is important to note here first that our construction is a set, so it may either contain an element (pair) or not, but it may never contain a duplicate pair; and second that there will in general be many covering sets for a given cross-product. For example, S''={ <a,a>,<a,b>,<b,c>} is also such that then *proj*$_1$(S')={a,b} and *proj*$_2$(S) = {a,b,c}, so S" also covers the full cross-product.

Given the notion covering set, the covering set difference d_{CS} between two potentially multi-valued data cells $d_{CS}(A, B)$ is the mean distance of the sum of the pair distances in the minimal cost covering set:

$$d_{CS}(A, B) \ \stackrel{\text{def}}{=}\ \frac{1}{|C|} \min_{C \in A \times B} d(C)$$

where we assume that the distance (cost) function d in use is simply lifted to sets in an obvious way, so that $d(C) = \sum_{p \in C} d(p)$, that we use the minimum cost cover (*min*), and finally, that $|C|$, as usual, denotes the cardinality of the (covering) set.

Nerbonne & Kleiweg (2003) then suggest that the minimal cost covering set be used to calculate the cost of the comparison of the multi-valued cells. It is not required that supplementary cost functions such as an inverse frequency weighting be included, but a

cost function might be included. For simple comparisons of categorical values, the cost function is simply zero for identical elements and one for non-identical ones.

The puzzling case of $A = \{a, b\}$ and $B = \{b, a\}$ is no longer problematic, since the minimal cost cover will simply be $C = \{< a, a >, < b, b >\}$, and $d_{CS}(C) = 0$.

2.2.2. A flaw in the definition

The problem is hinted at above, when we noted that there may be many covering sets. This would not necessarily lead to problems were it not the case that the covers may be of different cardinalities. Once we have covers of different cardinalities, then the mean differences will normally differ, so that we no longer assign a unique value to the difference between the multiply valued cells.

We illustrate the difficulty with an example from the original paper (Nerbonne & Kleiweg 2003):

> […] given $A = \{a, b, c\}$, $B = \{a, c, d\}$, then $C = \{< a, a >, < b, d >, < c, c >\}$ covers $A \times B$, even though $|C| = 3$, while $|A \times B| = 9$. Since $d(a, a) = d(c, c) = 0$, $d_{CS}(A, B) = d(b, d)/3$ […] (Nerbonne & Kleiweg, 2003: 349) [where we've added a subscript to the overall distance function d_{CS}, to keep it distinct from others we're considering].

This is right as far as it goes, and the cover C is indeed minimal in cost. The problem arises in considering other, equally minimally costing covers which may be greater in size. For example, $C' = \{< a, a >, < a, d >, < b, c >, < c, c >\}$ and is minimal given a cost function d such that $d(a, d) = d(b, c) = d(b, d)/2$. In this case $d(C') = d(a, d) + d(b, c)$, or $2d(b, d)/2$, i.e. $d(b, d)$. But now we're dealing with a 4-element cover, so the overall multivalued distance is $d(b, d)/|C|$, or $d(b, d)/4$. This demonstrates that the definition of d_{CS} fails to denote uniquely and is therefore flawed.

A similar example may readily arise in the case of string-valued attributes. Let $A = \{aaa, bcd, ecd\}$, $B = \{aaa, ecd, baa\}$. If we then calculate the differences between the strings using (normalized) Levenshtein distance (or the unnormalized variant), we note

that $LD\ (aaa, aaa) = 0\ =\ LD(ecd, ecd)$, while $LD(aaa, baa) = LD(bad, baa) = LD(bcd, ecd) = \frac{1}{3}$, $LD(bcd, baa) = \frac{2}{3}$, and $LD(aaa, ecd) = 1$. We then obtain the cover sets $C = \{< aaa, aaa >, < bcd, baa >, < ecd, ecd >\}$, but also $C' = \{< aaa, aaa >, < aaa, baa >, < bcd, ecd >, < bcd, bcd >\}$. The sums of the distances of the pairs in these two minimal covers are then $d(C) = d(baa, bcd) = \frac{2}{3}$ and $d(C') = d(aaa, baa) + d(bcd, ecd) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$. Both are minimal cost covers, costing 2/3 Levensthein units, but they differ in size, since $|C| = 3$, and $|C'| = 4$, so that there is no unique mean we can assign as the $d_{CS}\ (A, B)$.

The problem lies in looking for pairs whose distance is minimal and collecting them into a set without putting restrictions on the set size. We might consider trying to identify minimally sized minimal cost sets, but we propose that Bilbao distance (below) is simpler.

2.2.3 Other remarks

There is a second serious problem with the procedure sketched in this section, namely its computational complexity, as was noted in Nerbonne (2017). Examining all the potential subsets of the cross-product would mean examining all the subsets of AXB. There are |AXB| elements in the cross-product, so there are $2^{|AXB|}$ subsets. This is a large number in any case, e.g. if A has three elements and B five, then there are 2^{15} (> 32,000) subsets to examine. What is worse is that the number rises very steeply as A and/or B grows. Existing implementations have therefore been content to seek likely covers using some heuristics.

A third, more mathematical worry is that we should prefer a measure that satisfies the distance axioms, (zero distance between identical elements, symmetry and the "triangle inequality"),[3] and we have no proof that our "covering set" approach indeed yields a true distance measure.

[3] Wikipedia attributes the familiar metrical space axioms (Giles & Giles 1987: 1) to Maurice Frechet, https://en.wikipedia.org/wiki/Metric_space (January 1, 2020).

3. Bilbao distance

We introduce here Bilbao distance, designed to calculate the distance between two sets of strings or multiple (categorical) responses, $d_B(A, B)$, which can be presented in a compact way using the following formula:

$$d_B(A,B) = \frac{\sum_{i=1}^{|A|} \min_{b_j \in B} d(a_{i,} b_j) + \sum_{j=1}^{|B|} \min_{a_1 \in A} d(a_{i,} b_j)}{|A| + |B|}$$

This formula computes for each a_i in A a minimal distance with respect to the elements b_j in the other set B (i.e., for each a_i in A it computes a minimal distance with respect to the elements in the second set equal to $\min_{b_j \in B} d(a_{i,} b_j)$ analogously, it computes for each b_j in B a minimal distance respective to the elements a_i in A, the other set (i.e. for each in b_j in B it computes the minimal distance with respect to all a_i in A, equal to $\min_{a_i \in A} d(a_{i,} b_j)$), thus creating a *list* of $|A| + |B|$ minimal distances, whose mean will be the overall distance between A and B, denoted as above as $d_B(A, B)$.

Note that we proceed *not* from a *set* P of ordered pairs of elements whose minimal distance is summed, $d(C)$, as in the covering set construction above (in a set, pairs cannot be repeated), but rather directly on a *list* of minimal distances (in a list, distances can be repeated), so that, now, if a distance between a_i in A and b_j in B, denoted $d(a_i b_j)$, is the minimal distance for the string a_i with respect to every b_j in B (i.e. it holds that $\forall b_{j' \neq j}\ d(a_i, b_j) \leq d(a_i, b_{j,})$) and vice versa, i.e., $d(a_i b_j)$ is also the minimal distance for the value b_j with respect to every a_i in A (i.e. it also holds that $\forall a_{i' \neq i}\ d(a_i, b_j) \leq d(a_{i,}, b_j)$, then this distance $d(a_{i,} b_j)$ will be counted twice (not only once) when finally computing the mean between minimal distances (as said: $|A| + |B|$ minimal distances).

As an illustration, consider again the example provided by Nerbonne & Kleiweg (2003: 349), discussed above, where $A = \{a, b, c\},\ B = \{a, c, d\}$. We apply the simplest distance measure to categorical data, so that identical elements are zero distance from each other, and different elements are at a distance of 1. More formally $d(s_1, s_2) = 1 \leftrightarrow$

$s_1 \neq s_2$, and $d(s_1, s_2) = 0$ otherwise, i.e. when $(s_1 = s_2)$. So $d(a, a) = d(c, c) = 0$, and $d(b, d) = 1 \neq 0$ (because they are different strings).

To obtain the Bilbao distance between A and B, we first collect (and sum) the distances from each element in A to the element closest to it in B, i.e. $\sum_{i=1}^{k} \min_{b_j \in B} d(a_i, b_j)$. $A = \{a, b, c\}$, and we seek the elements in B closest to each of these in turn. The closest element to a in B, is just a, and $d(a, a) = 0$. Similarly for c, which is likewise found in both sets. But for $b \in A$, there is no matching element in B. No matter what we choose, the minimal distance will be $d(b, x) = 1$. We may conclude that $\sum_{i=1}^{3} \min_{b_j \in B} d(a_i, b_j) = 1$. A similar examination of all the elements of B, indicates that $\sum_{j=1}^{3} \min_{a_1 \in A} d(a_i, b_j) = 1$ because here, too, two of B's elements, namely a and c, are found in A, contributing zero to distance, while d has no match, so that its distance to each element in A is 1.

The Bilbao distance is just the sum of these contributions from the two sets, normalized over the sum of the sizes of the sets:

$$d_B(A, B) = \frac{\sum_{i=1}^{3} \min_{b_j \in B} d(a_i, b_j) + \sum_{j=1}^{3} \min_{a_1 \in A} d(a_i, b_j)}{|A| + |B|} = \frac{1 + 1}{3 + 3} = \frac{1}{3}$$

In addition to categorical values, we can examine sets of strings to illustrate how Bilbao distance works. We examine the case that was problematic for the covering sets approach. Let $A = \{aaa, bcd, ecd\}$, $B = \{aaa, ecd, baa\}$. The Levenshtein distances for all the string pairs are given above, in Sec. 2.2.2, i.e. as $LD(aaa, aaa) = 0 = LD(ecd, ecd)$, while $LD(aaa, baa) = LD(bcd, bce) = \frac{1}{3}$, $LD(bcd, baa) = \frac{2}{3}$, and $LD(aaa, ecd) = 1$. This leads us to gauge the distance between the two strings sets as follows:

$$d_B(A,B) = \frac{\sum_{i=1}^{3} \min_{b_j \in B} d(a_i,b_j) + \sum_{j=1}^{3} \min_{a_1 \in A} d(a_i,b_j)}{|A| + |B|} = \frac{((0 + \frac{1}{3} + 0) + (0 + 0 + \frac{1}{3}))}{3 + 3} = \frac{\frac{2}{3}}{6} = \frac{1}{9}$$

Just as in the more general example given above:

$$d_B(A,B) = \frac{LD(bcd,ecd) + LD(baa,aaa)}{6} = \frac{\frac{2}{3}}{6} = \frac{1}{9}$$

3.1 Some remarks on Bilbao distance

The Bilbao distance gives the same proportional weight to every string in A and B in order to obtain the total distance between A and B. This weight, $1/(|A| + |B|)$, is always inversely proportional to the total number of responses.

Bilbao distance also does not make the mistake of simply taking the mean of the cross-product, since $d_B(\{a,b\},\{b,a\}) = 0$, just as it should be. And it is computationally fairly efficient, considering each $a_i \in A$ with respect to each $b_j \in B$, and vice versa, making therefore $|A| \times |B|$comparisons of values, but twice, once for A, and once for B. So Bilbao distance calculation is a polynomial-time algorithm, a clear improvement over the exponential-time covering-set construction. One might speed the calculations up a bit using memoization techniques, but we have not pursued that.

3.2 An application to data

We consider it extremely interesting to add some examples with real linguistic data and see how the linguistic distance is computed using Bilbao distance, d_B. For this, responses collected in some localities referring to the concept of 'weather' (EHHA I: map 222) have been taken (localities in capital letters and answers in italics):

a) Distance between localities with only one answer each

Without any doubt, the distance between localities that have the same answers is 0, just as the distance between localities that have different answers is 1:

One locality with one answer (ELORRIO: *egualdi)* and a second locality with a different answer (LEMOIZ: *denpora).*

d_B (ELORRIO, LEMOIZ) = LD(*egualdi*, *denpora*) = 1, where LD is, as usual, relative Levenshtein distance.

b) Having one locality with more than one answer

One locality with only one word (LEMOIZ: *denpora)* and the second locality with two (DIMA: *egualdi*, *denpora)*, but one of these words is identical to the word of the first locality:

d_B (DIMA, LEMOIZ) = BD({*egualdi*, *denpora*},{*denpora*})= (0+0+LexD(*egualdi*, *denpora*))/3 = 1/3.

c) Having two localities both with MR-s:

If we consider two localities with two words in each: Dima *(egualdi, denbora)* and Legazpi *(egualdi, aro,* and taking into account that LD("egualdi", aro") = LD("denpora", aro") = 6/7), then:

d_B (LEGAZPI, DIMA) = d_B ({"egualdi", "aro"},{"egualdi", "denpora"}) = (LD("egualdi", "egualdi") + LD("aro", "egualidi") + LD("egualdi", "egualdi") + LD("denpora", "egualidi"))/4, which simplifies to (0 + LD("aro", "egualdi") + (0 + LD("denpora", "aro"))/4 = ((0 + 6/7 + 0 + 6/7)/4 = 12/28 = 3/7. Note that we might have chosen either "egualdi" or "denpora" as the element most similar to "aro", and similarly either "egualdi" or "aro" as most similar to "denpora". The distance does not change since both distances LD("aro", "egualdi") and LD("aro", "denpora") are equal to 6/7. Note, too, that d_B (LEGAZPI, DIMA) < ½ because LD("egualdi", aro") = LD("denpora", aro") = 6/7<1 (if these distances were =1, then d_B (LEGAZPI, DIMA) would be = ½).

4. Discussion and conclusion

The introduction, Sec. 1, demonstrated the need to deal with multiple values in dialectological work, showing *inter alia* that existing data collections often record multiple responses to queries. Sec. 2 reviewed previous attempts and sketched desiderata for a good treatment. Sec.2 also presented the most important negative result of this paper: the metric based on covering sets proposed by Nerbonne & Kleiweg (2003) is mathematically flawed since there are cases where it fails to assign a unique distance to a pair of multiply valued cells. We speculated about a potential repair, but did not develop the ideas in this paper. Instead we proposed Bilbao distance, which is weighted mean of the distances from each object in one set to its closest counterpart in the other.

The positive result is Bilbao distance, the new technique for assessing the size of the difference between two sets, which we presented and illustrated in Sec. 3. In sharp contrast to the covering-set construction, Bilbao distance eschews the construction of an optimal set in favor of simply summing up the distances from each element in both sets to the closest element in the other. We find this simplicity attractive.

The four desiderata we adduced earlier are realized by Bilbao distance. First, Bilbao distance is insensitive to the frequency with which a response occurs in the data, which is desirable since we usually cannot rely on counts in dialect survey data for genuine estimates of frequency because they are normally much too low. If we could, other techniques would suggest themselves, such as the multinomial regression found occasionally in sociolinguistic work (Paolillo 2018). We can apply Bilbao distance to the set any multiset reduces to without changing the result. Second, the distance between identical sets ought to be zero, just as for any distance metric. In particular $d_B(\{a,b\},\{b,a\}) = 0$.

Third, assuming the simplest categorical difference measure which assigns one to distinct elements and zero to identical elements, then we specified that the following should hold, and indeed it does: $d_B(\{a,b\},\{a\}) < d_B(\{a,b\},\{a,c\}) \leq 0.5$. In fact, $d_B(\{a,b\},\{a,c\}) = 0.5$, which seems reasonable, since a should not contribute anything to the distance. If $d_B(b,c) = 1$, as assumed, we intuitively see this comparison as contributing half of the difference to the set comparison. It is also intuitively appealing to

see that $d_B(\{a,b\},\{a\}) < d_B(\{a,b\},\{a,c\})$, since the differences are simply greater in the latter comparison. Fourth and finally Bilbao distance is eminently compatible with weightings that might be applied in comparison. This was demonstrated in the example in which Levenshtein distance was applied to the response strings, and other weightings (or cost functions) may likewise be applied directly.

We would also prefer that Bilbao distance be a true distance metric, i.e. that it assign only non-negative values, and zero to identical elements, that it be symmetric, and that it satisfy the triangle inequality. We cannot attempt a complete proof of this here, but we are optimistic. It can obviously assign only non-negative values assuming that this is what the underlying difference measure does (the function d in $\min_{b_j \in B} d(a_{i,} b_j)$ the definition). It will measure the difference between identical sets as zero assuming, again, that the underlying difference metric between the elements of these sets also assigns zero to identical elements. Its symmetry is obvious in the definition since it adds the partial differences with respect to each set, deriving the partial differences in the same way. So its symmetry follows from the commutativity of addition. The triangle inequality requires that $\forall A, B, C: D_B(A,B) \leq D_B(A,C) + D_B(C,B)$, and its status is more difficult to determine. Since the contribution of each item a_i in A is $\min_{b_j \in B} d(a_{i,} b_j)$, it would seem that, for there to be C such that $D_B(A,C) + D_B(C,B) < D_B(A,B)$, there would have to be $c_k, c_{k'} \in C$ such that $d(a_{i,} c_k) + d(c_{k'}, \min/a_i\,(b_j)) < \min_{b_j \in B} d(a_{i,} b_j)$, where $\min/a_i\,(b_j)$ is the b_j with the least distance to a_i. If $c_k = c_{k'}$ then the underlying difference metric would no longer satisfy the triangle inequality, so we may rule that possibility out, but we clearly need to keep in mind that the d_B will use the minimum c_k, i.e. $\min_{c_k \in C} d(a_{i,} c_k)$, and similarly in comparing sets B and C. The combinatorics can become complicated, leaving us unsure of this property.

In sum we propose that Bilbao distance be used to gauge the differences between cells of multiple values with low frequency. Future work should clearly include empirical application and attention to the unanswered mathematical question of the status of Bilbao distance with respect to the axioms of metric spaces.

References

ALEWIJNSE, Bart, John NERBONNE, Loke VAN DER VEEN & Franz MANNI (2007) "A computational analysis of Gabon varieties", in *Proceedings of the RANLP workshop on computational phonology*, Borovetz: Bulgaria: Recent Advances in Natural Language Processing Conference, 3-12.

ALLIÈRES, Jacques (1992) "La place de la variation synchronique ponctuelle dans les monographies dialectales et la géolinguistique", in G. Aurrekoetxea & X. Videgain (eds.), *Proceedings of International Congress on Dialectology,* Bilbao: Euskaltzaindia, 179-196.

AURREKOETXEA, Gotzon (2018) "Hizkuntza bariazioa eta erantzun anitzak" [Linguistic variation and multiple responses], in A. Etxebarria, A. Iglesias, H. Legarra & A. Romero (eds.), *Traineru bete lagun: Iñaki Gaminde omenduz,* Bilbao: UPV/EHU, 171-189.

BLANCQUAERT, Edgard & Willem PEÉ (eds.) (1925-1982) *Reeks Nederlandse dialect-atlassen*, Antwerpen: De Sikkel.

CHAMBERS, J. K. & Peter TRUDGILL (1998, [1]1978) *Dialectology*, Cambridge: Cambridge University Press.

EHHA: Euskaltzaindia (2010-2019) *Euskararen Herri Hizkeren Atlasa-EHHA,* vol. I-X, Bilbao: Euskaltzaindia.

FRANCO, Karlien (2017) *Concept features and lexical diversity. A dialectological case study on the relationship between meaning and variation*, PhD Diss., Catholic University of Leuven.

GILES, John R., and John Robilliard GILES (1987) *Introduction to the analysis of metric spaces*. Vol. 3, Cambridge: Cambridge University Press.

GOEBL, Hans (1984) *Dialektometrische Studien. Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF*. 3 Bände, Tübingen: Niemeyer.

GOEBL, Hans (1997) "Some dendrographic classifications of the data of CLAE 1 and CLAE 2", in W. Viereck & H. Ramisch (eds.), *The computer-developed linguistic atlas of England*, App. 9, Tübingen: Max Niemeyer, 23-32.

GOEBL, Hans (2010) "Dialectometry: Theoretical prerequisites, practical problems, and concrete applications (mainly with examples drawn from the *Atlas Linguistique de La France* (1902–1910)", *Dialectologia*. Special Issue I, 63-77.
<http://www.publicacions.ub.edu/revistes/dialectologiaSP2010/>

HEERINGA, Wilbert (2001) "De selectie en digitalisatie van dialecten en woorden uit de Reeks Nederlandse Dialectatlassen", *TABU: Bulletin voor taalwetenschap,* 31.1/2, 61-103.

HEERINGA, Wilbert, John NERBONNE & Peter KLEIWEG (2002) "Validating dialect comparison methods", in W. Gaul & G. Ritter (eds.), *Classification, Automation, and New Media*. *Proc. 24th Conf. Gesellschaft für Klassifikation*, Berlin: Springer, 445-452.

HINSKENS, Frans (2018) "Dialectology and Formal Linguistic Theory: The Blind Man and the Lame", in Ch. Boberg, J. Nerbonne & D. Watt (eds.), *The Handbook of Dialectology,* Boston: Wiley Blackwell, 88-105.

HOMBERT, Jean-Marie (1990) "Atlas linguistique du Gabon", *Revue Gabonaise des Sciences de l'Homme,* 2, 37-42.

HONEYBONE, Patrick (2011) "Variation and linguistic theory", in W. Maguire & A. McMahon (eds.), *Analysing variation in English*, Cambridge: Cambridge University Press, 151-177.

KRETZSCHMAR Jr, William A., Virginia G. MCDAVID, Theodore K. LERUD & Ellen JOHNSON (eds.) (1994) *Handbook of the Linguistic Atlas of the Middle and South Atlantic States*, Chicago: University of Chicago Press.

LABOV, William (1969) "Contraction, deletion and inherent variability of the English copula", *Language,* 45, 715-762.

MANNI, Franz & John NERBONNE (to appear) "Linguistic Diversity and Human Migrations in Gabon", in L. Muñoz & M. Crawford (eds.), *Human migration: Bio-cultural perspectives*, Oxford: OUP. Preprint available at www.let.rug.nl/nerbonne/papers.html

MANNING, Christopher & Henrich SCHÜTZE (1999) *Foundations of statistical natural language processing*, Cambridge: MIT Press.

NERBONNE, John (2003) "Linguistic variation and computation", *Proceedings of the 10th EACL conference*, Shroudsburg, Penn.: Association for Computational Linguistics, 3-10.

NERBONNE, John (2017) "Respecting local variation", in A. Iglesias & A. Ensunza (eds.), *Gotzon Aurrekoetxea lagunarterik hara,* Bilbao: UPV/EHU, 13-24.

NERBONNE, John & Wilbert HEERINGA (2010) "Measuring dialect differences", in J.-E. Schmidt & P. Auer (eds.), *Language and Space: Theories and Methods,* Berlin: Mouton De Gruyter, 550-566.

NERBONNE, John & Peter KLEIWEG (2003) "Lexical distance in LAMSAS", *Computers and the Humanities,* 37, 339-357.

PAOLILLO, John C. (2018) "Logistic Regression Analysis of Linguistic Data", in Ch. Boberg, J. Nerbonne & D. Watt (eds.), *The Handbook of Dialectology*, Boston: Wiley, 384-399.

PEREA, Maria-Pilar (2009) “La dialectometría y su aplicación en el estudio de las variedades dialectales del catalán”, *Revista de Filoloxía Asturiana,* 9-10, 109-130.

PROKIĆ, Jelena, Çağri ÇÖLTEKIN & John NERBONNE (2012) "Detecting shibboleths", *Proceedings of the EACL Joint Workshop LINGVIS & UNCLH*, Shroudsburg, PA: Association for Computational Linguistics, 72-80.

RAVIER, X. (1978-1993) *Atlas linguistique et ethnographique du Languedoc Occidental,* Paris: CNRS.

RUMPF, Jonas, Simon PICKL, Stephan ELSPASS, Werner KÖNIG & Volker SCHMIDT (2009) “Structural analysis of dialect maps using methods from spatial statistics”, *Zeitschrift für Dialektologie und Linguistik,* 76/3, 280-308.

SÉGUY, Jean (1973) “La dialectométrie dans l'Atlas linguistique de la Gascogne”, *Revue de Linguistique Romane,* 37, 1-24.

SOUSA, Xulio (2017) “From field notebooks to automatic mapping: the ‘Atlas Lingüístico Galego’ database”, *DiG,* 25, 1-22.

TAELDEMAN Johan & A. GOEMAN (1996) “Fonologie en morfologie van de Nederlandse dialecten: Een nieuwe materiaalverzameling en twee nieuwe atlasprojecten”, *Taal en Tongval*, 48, 38-59.

VENY, Joan & Lídia PONS I GRIERA (dir.) (2001-2018) *Atles lingüístic del domini Català.* Vol. 1-9, Barcelona: Institut d’Estudis Catalans.

VIDEGAIN, Xarles & Gotzon AURREKOETXEA (2015) “Lehen kolpearen teoria eta EHHA”, in G. Aurrekoetxea, A. Etxebarria, A. Romero (eds.), *Linguistic variation in the Basque language and education-I*, Bilbao: UPV/EHU, 114-126.

WIELING, Martijn, Wilbert HEERINGA & John NERBONNE (2007) “An aggregate analysis of pronunciation in the Goeman-Taeldeman-Van Reenen-Project data”, *Taal en Tongval,* 59.1, 84-116.

WIELING, Martijn & John NERBONNE (2015) "Advances in dialectometry”, *Annual Review of Linguistics,* 1.1, 243-264.