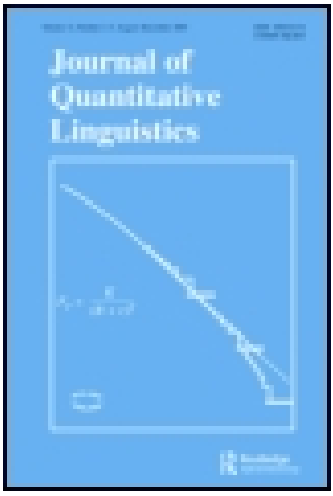

Journal of Quantitative Linguistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/njql20

Counting Languages in Dialect Continua Using the Criterion of Mutual Intelligibility*

Harald Hammarström [a]
[a] Chalmers University of Technology , Gothenburg
Published online: 16 Jan 2008.

To cite this article: Harald Hammarström (2008) Counting Languages in Dialect Continua Using the Criterion of Mutual Intelligibility*, Journal of Quantitative Linguistics, 15:1, 34-45, DOI: 10.1080/09296170701794278

To link to this article: http://dx.doi.org/10.1080/09296170701794278

Journal of Quantitative Linguistics
2008, Volume 15, Number 1, pp. 34 – 45
DOI: 10.1080/09296170701794278

Routledge
Taylor & Francis Group

Counting Languages in Dialect Continua Using the Criterion of Mutual Intelligibility*

Harald Hammarström
Chalmers University of Technology, Gothenburg

ABSTRACT

This paper shows how it is possible to count languages vs. dialects if, for every pair of varieties, we are given whether they are mutually intelligible or not. The method is to divide the varieties into a minimum number of internally mutually intelligible groups where each group counts as one language. Expressed in terms of graphs (as in discrete mathematics), the method is even easier understood as: applying graph-colouring to a graph over varieties with the intelligibility interrelationships as edges. Graph colouring is already mathematically well-understood and we can easily prove properties intuitively associated with the concepts language and dialect, and remove any fears that these concepts should lead to inconsistencies. The presentation requires only a minimal acquaintance with sets, combinatorics and graphs.

1. INTRODUCTION

In trying to answer the question "How many languages are there in the world?" linguists have had a hard time coming up with a satisfactory answer. Even when explicitly disregarding non-linguistic criteria (such as ethno-socio-economico-politico-cultural ones), they say that defining languages by the mutual intelligibility criterion (MI) is not possible (e.g. Anderson, 2005).

Firstly, mutual intelligibility is not a strict yes/no distinction but a matter of degree. Subsumed hereunder are also cases where the degree of intelligibility is not enough to enable communication immediately, but

*Address correspondence to: Harald Hammerström, Department of Computing Science, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden.
E-mail: harald2@cs.chalmers.se

high enough to enable communication after, say, only a few days of exposure, such as among the Mekeo languages (Jones, 1998, p. 19). Also, cases may exist where intelligibility is not symmetric, i.e. *A* understands *B* but not vice versa, although I have yet to see a genuine well-documented example.[1]

Secondly, even if it were simplified into being yes/no and symmetric, counting languages by the MI would lead to contradictions in dialect-chain situations. For example, if *A* is MI with *B*, *B* is MI with *C* but *A* is not MI with *C* – a completely realistic situation – then setting *A* and *B* to be the same language and *B* and *C* as the same language is contradictory because *A* and *C* are not the same language by the MI criterion.

In this paper we will show that the second objection is premature. There is a perfectly consistent way to count languages using a symmetric strict yes/no mutual intelligibility criterion that preserves intuitive properties about languages and numbers of languages. Linguists seem to have so far failed to appreciate this,[2] as the following selection of quotations exemplify:

> Such situation are referred to by linguists as ''dialect chains'', and they result in sometimes arbitrary decisions being made as to how many languages are involved.
>
> (Lynch & Crowley, 2001, p. 2)

> The criterion that ''technically . . . mutually intelligible forms of speech are known as **dialects**, and [that] the term **language** is used for mutually unintelligible forms of speech'' (Lehmann, 1973, p. 33), does not apply satisfactorily to such situations as the Chaga continuum from Siha to Usseri.
>
> (Polomé, 1980, p. 3)

[1]Of course I can think of examples where *A* and *B* are closely related and speakers of *A* tend to understand *B* but not the other way round. For the sake of an example take Jamaican Creole and Oxford English. But in most (all?) such cases this is because the *A* speakers have been exposed to *B* a lot more, and not purely because of their knowledge of *A*. I see no reason to differentiate this situation from that where *A* are *B* aren't closely related, and speakers of *A* know *B* as well, but not vice versa.

[2]See, however, Hockett (1958, pp. 321 – 330) for an embryo to the approach taken in this paper (whose views recollected by, for example, Heine & Köhler (1981, pp. 1 – 3). Note also that the matter is not discussed in the most recent encyclopaedia entry on dialect chains (Heap, 2006).

> A common situation is a string of similar varieties, in which the speakers of variety A understand those of C, and so on, but the speakers of A do not understand the variety at the other end of the continuum, or even those part way along. Even if we can define 'understand', where is the divide between language and dialect in this situation?
>
> (Heine & Nurse, 2000, p. 2)

> In some cases, the intelligibility criterion actually leads to contradictory results, namely when we have a dialect chain, i.e. a string of dialects such that the adjacent dialects are readily mutually intelligible, but dialects from the far ends of the chain are not mutually intelligible. A good illustration of this is the Dutch-German dialect complex. One could start from the far south of the German-speaking area and move to the far west of the Dutch-speaking area without encountering any sharp boundary across which mutual intelligibility is broken; but the two end points of this chain are speech varieties so different from one another that there is no mutual intelligibility possible. If one takes a simplified dialect chain A – B – C, where A and B are mutually intelligible, as are B and C, but A and C are mutually unintelligible, then one arrives at the contradictory result that A and B are dialects of the same language, B and C are dialects of the same language, but A and C are different languages. There is in fact no way of resolving this contradiction if we maintain the traditional strict difference between language and dialects, and what such examples show is that this is not an all-or-nothing distinction, but rather a continuum. In this sense, it is impossible to answer the question how many languages are spoken in the world.
>
> (Comrie, 1987, p. 3)

On the latter quote, a few clarifying remarks are in order: Comrie is discussing the definition that runs "two varieties are the same language **if and only if** they are mutually intelligible".[3] I am not denying that this definition leads to a contradiction – I am saying that other intuitively

[3]Comrie affirms in a personal email (9 September 2005) that the quoted paragraph concerns only this particular definition, and the statements therein that may look as if they quantify also over other intuitive definitions based on the MI, should not be so interpreted.

acceptable definitions based (solely) on the MI have been ignored. In particular, I will present a definition whereby it is still true that "if two varieties are the same language then they are mutually intelligible" but the converse "if two varieties are mutually intelligible then they are of the same language" does not have to be true. The spirit of the latter is instead rendered by a slightly more relaxed requirement that says that the number of languages should not be unnecessarily multiplied. The definition is given full formal treatment below.

2. COUNTING LANGUAGES

The task is to decide, for a finite set X of speech varieties, how many languages there are using only a binary symmetric strict yes/no relation of mutual intelligibility (henceforth MI). For ease of presentation we shall model the situation as there being n speakers each speaking exactly one variety. It will be seen that the method is really indifferent to the distribution of varieties over people, names or any other grouping, so there is no loss of generality. Speakers will be denoted by capital letters, e.g. A, B, C. Thus let $X = \{A, B, C, \ldots\}$ be any finite set of speakers.

2.1 Definition

Definition 1. *The number of languages in X is the least k such that one can partition X into k blocks such that all members within a block understand each other.*

A partition of a set X into blocks is simply a division of the members of X into disjoint non-empty groups that exhaust X. So if say, $X = \{A, B, C\}$, we can partition it into:

1. One block: $\{A, B, C\}$.
2. Two blocks: there are exactly three possibilities $\{A, B\},\{C\}$ or $\{A\},\{B, C\}$ or $\{A, C\},\{B\}$.
3. Three blocks: $\{A\},\{B\},\{C\}$.

Clearly, the number of blocks in a partition ranges between 1 and the number of members of the set (also known as the cardinality of the set).

Now let us say $X = \{A, B, C\}$ depicts the classic dialect chain situation where A and B are MI, B and C are MI, but A and C are not MI.

The partition into one block does not satisfy the requirement of the definition, since A and C that are in the same block, do not understand each other. Of the three partitions into $k = 2$ blocks, two of them satisfy the definition: $\{A, B\}$, $\{C\}$ is ok because A and B are MI; $\{A\}$, $\{B, C\}$ is ok because B and C are MI. The partition into three blocks also trivially satisfies the condition that no pair within a block should be mutually unintelligible, but $k = 3$ is not minimal. Thus the number of languages in the example is 2, and we can immediately observe a curious feature of the definition: the number of languages k is unique, but there may be several satisfying partitions into k blocks.

2.2 Properties

It should be obvious that the definition is well-behaved in the sense it yields a unique number of languages k (for example, one way to arrive at the number is to just try out all partitions of the given set X). It should also be clear that a partition defines an assignment of speech varieties into languages such that A and B belong to the same language if and only if they belong to the same block. But what about properties of partitions that satisfy the minimal k and the requirement of inside-block intelligibility? Intuitively, if blocks are to be identified with languages, one would expect the following two properties to hold:

Property 1. *All those who speak the same language speak varieties which are mutually intelligible.*

Property 2. *There are no "superflous" languages, i.e. a person speaking exactly one variety of each of the languages can communicate with everyone, whereas someone speaking less than k varieties cannot communicate with everyone.*

That the first property holds for the given definition is immediate from the definition. Informally, the second property holds because otherwise k would not be minimal, as required. A more detailed proof, which involves a little more work, is given below in Section 3.3.

The definition and the properties emanating from it, is not specific to any particular type of language-variety landscape but extends to completely arbitrary constellations of language varieties and MI-interrelationships. It should actually be understood as an even more abstract counting method: given a set of objects and a symmetric,

non-reflexive, non-transitive ''is-different'' relation over them (here: mutual unintelligibility), what is the minimal number of blocks one can partition this set into such that all members within a block are not different to each other?

For pedagogical reasons we shall now continue the presentation in terms of graphs.

3. FURTHER EXAMPLES AND PROPERTIES

Again, the task is to decide, for n speech varieties, how many languages there are using only a binary symmetric strict yes/no relation of mutual intelligibility.

3.1 Definition

Let the n speakers be vertices V of a graph[4] G. Let G have an edge between vertices A, $B \in V$ if and only if A and B do **not** speak mutually intelligible varieties.

Definition 2. *The number of languages is the smallest k such that one can colour the vertices of G with k colours such that no two vertices that share an edge have the same colour.*

This number is usually called the ''chromatic number'' of a graph G and is denoted $\chi(G)$ (Read, 1968).

3.2 Examples

Again, an example of a graph illustrating the most basic dialect-chain situation is shown in Figure 1. For G in Figure 1 the chromatic number is 2. It is not possible to colour the vertices A, B and C with only one colour because then A and C would get the same colour – violating the condition that vertices which share an edge should not have the same colour. It would be possible to colour the vertices with (exactly) three colours, one each, without violating the shared-edge-different-colour

[4]For readers not familiar with graphs, a graph can be thought of as a set of points (''vertices'') in a two-dimensional space and an arbitrary set of lines (''edges'') between pairs of points. More information can be found in any introductory book on discrete mathematics.

condition, but 3 is not the chromatic number because it is also possible to colour G with less, namely 2, colours. There are in fact two different ways to colour G with 2 colours (say red and green): 1. $\{A, B\}$ red, $\{C\}$ green; 2. $\{A\}$ red, $\{B, C\}$ green.

Another example, a four-member dialect chain, is shown in Figure 2. For G in Figure 2 the chromatic number is also 2. There is only one 2-colouring: A, B green and C, D red. (Clearly, one colour is not sufficient. However, if one tries two colours: A must have some colour, then C and D must have a different colour. Nothing prevents C and D from having the same colour, so they get the same colour. Now, only B remains which cannot be coloured by the colour of C and D, but it can have A's colour. All the choices were forced or colour-conservative so this is the only two-colour possibility.)

Lastly, a third slightly more complicated example is shown in Figure 3. The chromatic number of the graph in Figure 3 is 3. There are no less

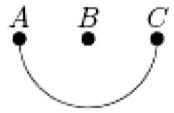

Fig. 1. Graph for (A, B), (B, C) are MI but (A, C) are not MI.

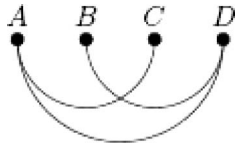

Fig. 2. Graph for (A, B), (B, C), (C, D) are MI but no other pairs are MI.

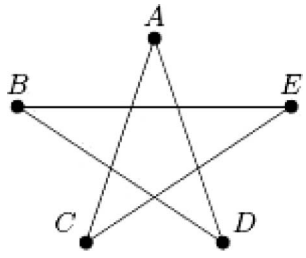

Fig. 3. Graph for (A, B), (B, C), (C, D), (D, E), (E, A) are MI but no other pairs are MI.

than five different minimal colourings: 1. {*A*} {*B*, *D*} {*C*, *E*}; 2. {*B*} {*A*, *D*} {*C*, *E*}; 3. {*A*, *D*} {*B*, *E*} {*C*}; 4. {*A*, *C*} {*B*, *E*} {*D*}; 5. {*A*, *C*} {*B*, *D*} {*E*}. This is perhaps more easily seen if the graph is redrawn (not changed) as a pentagon, i.e. by keeping *A* at the top, but putting *D* and *C* at the next level and *B* and *E* at the bottom level (as shown in Figure 4).

3.3 Properties

As may have been experienced by the reader, it is not trivial to calculate what the chromatic number is, even for a graph of quite moderate size. It should be clear to everyone though, that it is always possible to reach the answer by tediously enumerating and checking all possibilities.

There is a mathematically well-understood systematic method to calculate the chromatic number (and the number of minimal-size colourings), in terms of breaking down any graph into more tractable pieces. The interested reader may consult the excellent introduction by Read (1968).

The bad news is that finding the chromatic number is an NP-complete problem (Garey & Johnson, 1979). In layman terms, this implies that there is no known "smarter" method to find the chromatic number than to go through all the possible partitions of the vertices into blocks ("colours"). All known methods rely on the fact that it is "easy" to check whether a given colouring is ok, i.e. just to check if any edges violate the constraint, but still, in the worst case, we need to step through essentially all possible partitions. This is the bottleneck because the number of possible partitions of *n* vertices in blocks is exponential[5] in *n*. For

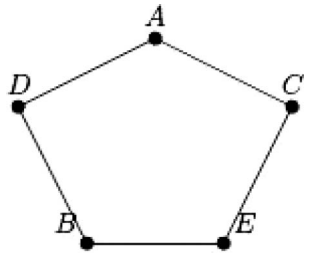

Fig. 4. Graph for (*A*, *B*), (*B*, *C*), (*C*, *D*), (*D*, *E*), (*E*, *A*) are MI but no other pairs are MI.

[5]In fact, it is $S(n, 1) + S(n, 2) + \ldots + S(n, n)$, where *S* denotes the Stirling numbers of the second kind. See, for example, Stanley (1997, p. 33) for more information.

instance, if $n = 20$ there are 51,724,158,235,372 partitions to consider. Such is the problem in its full generality. But of course, for specific cases, there may be symmetries and regularities which make the solution considerably more digestible.

As has been observed, there may be more than one minimal colouring. For any of one these minimal colourings, we can identify colours as languages. That is, let G again be the graph depicting the situation at hand and let $c_1, c_2, \ldots, c_k \neq \emptyset$ be a minimal colouring, thus satisfying $k = \chi(G)$ (the chromatic number of G), $\cup C_i = V$ and $C_i \cap C_j \neq \emptyset$ for $i \neq j$. Identifying languages as colours simply means that $c_1, c_2, \ldots, c_k$ form the k languages. Languages so defined have the following two crucial properties:

Property 1. *All those who speak the same language speak varieties which are mutually intelligible.*

Property 2. *There are no "superflous" languages, i.e. a person speaking exactly one variety of each of the languages can communicate with everyone, whereas someone speaking less than k varieties cannot communicate with everyone.*

Proof of Property 1: Assume there were a pair of non-MI varieties of the same language, i.e. assigned the same colour. Since they were not MI they would share an edge in the graph – contradicting that the colouring was legal in the first place.

Proof of Property 2: Say that one speaks $k' < k$ varieties $Z_1, Z_2, \ldots, Z_{k'}$. Form the k' groups one could communicate with using the respective variety: $C'_i = \{Y \in V | Y \text{ is MI with } Z_i\}$ for $1 \leq i \leq k'$. (If $C'_i \cap C'_j \neq \emptyset$ for some $i < j$ then remove the intersecting elements from (say) C'_i). If one could communicate with everyone using the Z_i varieties, then $\cup C'_i = V$. Since within each C'_i there are no edges between the members (they are MI by definition), $C'_1, C'_2, \ldots, C'_{k'}$ would yield a legal colouring of G – contradicting the minimality of k.

Restating, any minimal colouring of a (graph of) a language/dialect situation has the above two properties. Any count of the number languages other than "as many as the blocks of a minimal colouring" would in some way fail to satisfy one of the two properties about the

languages counted. And clearly, the two stated properties must be part of the intuitive understanding of what it means to be a language.

At this point, however, we still cannot give a finished count of the number languages of the world because:

- We still do not have an answer to *when* mutual intelligibility holds given two languages (cf. the first point in the introduction).
- Even if we did have a good (or arbitrary) method to decide when two varieties are mutually intelligible, we do not have complete knowledge of the speech varieties of the world. The best one-piece source on this matter is the *Ethnologue* (Gordon, 2005) but it does not provide (nor does it aim to) systematic detailed information on (any kind of) intelligibility between varieties.
- As alluded to above, even if we did have complete knowledge etc, the resulting graph for the world would have on the order of 6900 vertices, which might be intractable. Since most language varieties, unquestionably, aren't intelligible to each other, this graph would turn out to be quite easily handled, but it remains to see just how complex it does get.

REFERENCES

Anderson, S. R. (2005). *How many languages are there in the world?* Answer to a FAQ by the Linguistic Society of America, Washington, D.C. Retrieved 1 September 2005 from http://www.lsadc.org/pdf_files/howmany.pdf

Comrie, B. (Ed.) (1987). *The World's Major Languages*. London: Croom Helm.

Garey, M. R., & Johnson, D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. New York: W. H. Freeman.

Gordon Jr., R. G. (Ed.) (2005). *Ethnologue: Languages of the World,* 15th edition. Dallas: SIL International.

Heap, D. (2006). Dialect chains. In K. Brown (Ed.), *Encyclopedia of Language and Linguistics* (pp. 528–530). 2nd edition, Volume 3. Amsterdam: Elsevier.

Heine, B., & Köhler, O. (1981). *Linguistik - Ostafrika (Kenya, Uganda, Tanzania): Gliederung der Sprachen und Dialekte*. Afrika-Kartenwerk E 10. Berlin: Gebrüder Borntraeger.

Heine, B., & Nurse, D. (2000). Introduction. In B. Heine & D. Nurse (Eds), *African Languages: An Introduction* (pp. 1–10). Cambridge: Cambridge University Press.

Hockett, C. F. (1958). *A Course in Modern Linguistics*. Toronto: Macmillan.

Jones, A. A. (1998). *Towards a Lexicogrammar of Mekeo (An Austronesian Language of West Central Papua)*. Volume 138 of Pacific Linguistics: Series C. Canberra: Research School of Pacific and Asian Studies, Australian National University.

Lynch, J., & Crowley, T. (2001). *Languages of Vanuatu: A New Survey and Bibliography*. Volume 517 of Pacific Linguistics. Canberra: Research School of Pacific and Asian Studies, Australian National University.

Polomé, E. C. (1980). The languages of Tanzania. In E. C. Polomé & C. P. Hill (Eds), *Language in Tanzania* (pp. 1–25). Ford Foundation Language Surveys, Oxford University Press.

Read, R. C. (1968). An introduction to chromatic polynomials. *Journal of Combinatorial Theory*, *4*, 52–71.

Stanley, R. P. (1997). *Enumerative Combinatorics: Volume I*. Volume 49 of Cambridge Studies in Advanced Mathematics. Cambridge: Cambridge University Press.

APPENDIX: FAILING APPROACHES TO DEFINING LANGUAGES UNIQUELY

As seen in the previous section, the method for counting languages does not always uniquely determine "which" the languages are, it just says "how many" they are and the number of alternatives to which they are. It is tempting to look at ways to synthesize the alternatives into a unique language/dialect definition.

For example, it is readily seen in Figure 1 that in the two minimal colourings, A and C are never in the same colour, whereas B is once in the company of A and once in C. It is then tempting to define the languages as "the maximum-size set L of vertices that are never in the same colour in any minimal colouring". The rest of the vertices can then be thought of as dialects of the varieties to which they are MI. (A dialect can then be the dialect of several (distinct) languages, and even be a dialect of a dialect). In the example of Figure 1 this would beautifully synthesize the two minimal colourings to say that A and C are separate languages and B is a dialect of A as well as a dialect of C.

Unfortunately there are cases where this approach does not "work". In the graph of Figure 2, L would not be unique; either of $L = \{A, C\}, \{A, D\}, \{B, C\}, \{B, D\}$ has the maximum-size 2. (Here, of course, since there is only one minimal colouring, we can satisfactorily take AB and CD as the two languages and call A a dialect of AB, B another dialect of AB, and so on).

But, more seriously, in the case of Figure 4 (depicting the same situation as in Figure 1), there are five different maximum-size $L = \{A, C\}, \{A, D\}, \{B, D\}, \{B, E\}, \{C, E\}$, so $|L| \neq k$ which invalidates the idea. The graph of Figure 4 is the smallest graph where $|L| \neq k$. I don't believe there is a sensible way of uniquely defining languages in such graphs (that is, graphs which have odd-size circles but whose chromatic number is lower than the number of members of the circle). In the graph in

question, we are told that there are exactly three languages but all the vertices are symmetric so there seems to be no way to single out three of them or divide five nodes into three equal-size groups. If we wish to select three of the five to be languages and the other two dialects, there is no unarbitrary way to decide which go as languages since all five vertices are structurally indistinguishable. If we wish to divide the five into three groups, they would not all be of equal size and, again, there is no basis for putting one or the other vertex in the bigger (or smaller) group.

This might not just be a purely theoretical problem. It is conceivable that such a ''ring'' could be the correct state of affairs somewhere in the world, say, if a dialect continuum settled around a mountain and the two extremes of the chain meet and influence each other (for a couple of centuries) so that they become intelligible dialects.