Measuring languageness:
Fact-checking and debunking a few common myths

Mauro Tosco[1]
[1]University of Turin

Abstract – The article critically discusses a few common objections to an intrinsic, language-internal definition of what constitutes a 'language' (and, conversely, a 'dialect'). It argues that, contra postmodernism (1.), languages do exist, they can be counted and languageness can be measured independently and even notwithstanding the speakers' beliefs and ideologies (2.). It further refutes as unsound all the common criticisms to intelligibility as a tool in assessing languageness: while deviations from common-sense assessments may be expected but are not really of concern to science (3.1.), intelligibility asymmetries (3.2.), apparent infinite graduality (3.3.) and dialect chains (3.4.) are only partial problems to be solved empirically. On the contrary, intelligibility can and is routinely measured in different sciences, and, when applied to language, it tends to dovetail with other criteria, such as dialectometry and the counting of isoglosses (4.).

Keywords: languageness; measurability; dialects; postmodernism; intelligibility.

1. There are no languages (and nothing to be measured)

Postmodernism has had its sway in many areas of science, especially soft ones. Whole new fields have been born out of it. Its antipositivist philosophy goes hand in hand with its radical relativism and its dangerous antiscientific bias has been remarked many times. In a postmodernist perspective, languages are either a figment of imagination (Makoni 2005) or merely the result of political acts (Pennycook 2007). Language names are at best tags, and only endless variation and change is real. Language becomes a 'narrative', and narratives are a special target of postmodernism. Even if, as a postmodernist, you are not particularly keen on logical reasoning, once you get rid of the "pernicious myth" (Pennycook 2006: 67) of languages, you can't really delve too much into language rights and policy or, simply, language studies.

Not based upon any empirical evidence (empiricism being 'the' bogeyman), there is not really any myth to debunk here. You simply buy it or not.

It looks more interesting and maybe promising to take a closer look at linguistics' own problems with the very notion of 'language' (and related 'dialect'), and to discover that postmodernism has fed upon fertile grounds in language matters: stripped of its radical overtones, the notion that 'languages' and 'dialects' are basically social constructs and therefore can only be defined in terms of socio-political status and breadth of use, is so common currency in textbooks to be almost a platitude, almost on a par with the old jokes on languages with navies and dialects without them. This is the view qualified of *Ausbau*-centrism in Tamburelli and Tosco (2021): out of the two poles of Kloss' (1967) dichotomy *Ausbau* vs. *Abstand*, mainstream general linguistics has chosen to define languages alongside the dimension of *Ausbau*, i.e., their role as a "standardised tool of literary expression" (Kloss 1967: 69), and irrespective of their linguistic distance (i.e., *Abstand*). Exit *Abstand*: after *Ausbau*'s landslide victory, what remains is a discipline refusing to define the very entity from which it takes its name. But, as aptly remarked by Nunberg (1997: 675), if defining languages is not our pigeon, whose is it?

The alleged, generally unsound, reasons of such an attitude are discussed in the remainder of this article.

2. Languageness and self-identification

How can we define what is a language and what is not? Self-identification is an apparently good (and easy) solution; it runs roughly like this: "[I]f there is a social group that believes and acts as if a linguistic system is a language then it is one" (Fasold 2005: 698). In such a solution the burden of establishing languageness is placed squarely upon the speakers, and the linguists' role is to register their verdict.

Essentially, this is the same sociolinguistic definition of languageness based upon navies and armies and mentioned above. But with a convenient democratic twist: majority rule (: a social group's belief) takes the place of power (: navies, i.e., brute strength) in the establishment of truth.

But what is a 'group' or a 'community'? It seems to be, circularly, 'a group of people who holds the same opinion on X and acts accordingly'. As any 'community' (at least of humans; other species may behave differently but the question is immaterial here) is rarely (if ever) unanimous in perceptions and judgments, any reference to a community's beliefs is at most statistically true. As one cannot imagine why problems pertaining to language would be any different than any other venue of human experience, a community's judgment on languages is therefore a statistical truth, too.

Interestingly, the above does not change even when such beliefs and judgments are cloaked in terms of 'democracy' (which is supposedly based upon the decisions of a majority of the group's members).

This would further imply that languages are social constructs ('what is perceived by a community to be a language'): we are back to square one and postmodernism. Now, the study of social construct is certainly an interesting and useful enterprise and may even be scientific under a fairly liberal understanding of what constitutes 'science'. What is not and cannot be is equivalent to the study of taxonomically independently identifiable entities. Nor are the two mutually exclusive. Just as the evidence gathered from the study of folk taxonomies does not impinge on the validity of Linnean classification, the study of what community **α** thinks of language X does not make X a language (nor it makes **α** right).

Ironically, Pennycook (2007: 91) himself comes in support here: "majority belief doesn't tell us anything about the existence of what is believed in".

There seems to be no way out than to firmly reject the speakers' attitudes, ideologies and beliefs, and to place the burden of establishing languageness squarely on the shoulders of the linguists' community, however weak and unprepared to the task they may be.

3. Intelligibility and its enemies: debunking a few myths

If to understand and make oneself understood is pivotal to the layman's definition of what it means 'to speak a language' and be part of a language group, it is paradoxical that so much effort has been spent on the part of so many linguists in order to show that intelligibility cannot be proven – nor therefore measured. As such, we are told, it is not even a linguistic problem. Period.

Many a claim that intelligibility cannot be measured has been debunked by Tamburelli (2014) and the interested reader is referred to his work for more details. Here I will briefly review and expand on a few points.

3.1. "Too many languages"

Discussing the very definition of language, Comrie (1987: 2) observes:

> [I]f two speech varieties are mutually intelligible, they are different dialects of the same language, but if they are mutually unintelligible, they are different languages. But if applied to the languages of Europe, this criterion would radically alter our assessment of what the different languages of Europe are: the most northern dialects and the most southern dialects (in the traditional sense) of German are mutually unintelligible, while dialects of German spoken close to the Dutch border are mutually intelligible with dialects of Dutch spoken just across the border. In fact, our criterion for whether a dialect is Dutch or German relates in large measure to social factors – is the dialect spoken in an area where Dutch is the standard language or where German is the standard language? By the same criterion, the three nuclear Scandinavian languages (in the traditional sense), Danish, Norwegian and Swedish, would turn out to be dialects of one language, given their mutual intelligibility. While this criterion is often applied to non-European languages (so that nowadays linguists talk of the Chinese languages rather than the Chinese dialects, given the mutual unintelligibility of, for instance, Mandarin and Cantonese), it seems unfair that it should not be applied consistently to European languages as well.

Comrie's analysis intersects here with the problem of dialect continua to be discussed in 3.4. Let us discuss instead his conclusion that "this criterion would radically alter our assessment of what the different languages of Europe are"; it tallies neatly with Trudgill's (2000: 4):

> [W]e could say that if two speakers cannot understand one another, then they are speaking different languages. Similarly, if they can understand each other, we could say that they are speaking dialects of the same language. Clearly, however, this would lead to some rather strange results in the case of Dutch and German, and indeed in many other cases.

Practicality and, strangely enough, 'fairness' and 'strangeness' have here the upper hand. As such, there is no real answer and, with nothing to be proven, nothing can be *dis*proven: maybe our assessments would change and maybe not (probably not much, as our discussion is bound to show); it remains instead true that many would think of this as 'unfair' and 'strange' (after all, Italian undergraduates are often shocked when learning that, according to the most reliable estimates, more than 30 indigenous languages are spoken in a country that fought and is still fighting so hard to become monoglottic).

Maybe, in the end we will indeed come up with "too many languages" and "strange, unfair results". So what?

3.2. "Intelligibility may be asymmetric"

That intelligibility may be asymmetric (at a social level, of course) is an oft-repeated argument against its possible use in measuring languageness and the distance between dialects and languages in particular. Differences in reciprocal intelligibility between speakers of Spanish and Portuguese, or of different Arabic dialects, are often invoked. There is striking lack of factual data offered to back such assertions: once again, anecdotical evidence takes the place of carefully designed research, scientific hypotheses, experiments, and figures.

But, again, so what? Intelligibility is often asymmetric specifically in the case of minority languages, where all the speakers of the minor group are bilingual in the bigger group's idiom. Quite often, minority language speakers are willy-nilly adopting the language of the majority and often get even more conversant in the majority language than in their ancestral idiom. It is

just the – often not so long – road to language shift and language death. In all these cases the majority language speakers can forgo the pain to learn the other language, and in all these cases intelligibility is indeed asymmetric. The oft-mentioned instances of asymmetric intelligibility are not that different in kind and essence from the common, everyday experience of second language speakers and learners of language A vs. A's monolingual speakers and, as Wolff (1959) pointed out, they often often boil down to the (passive) acquired knowledge of a variety, or longer exposure to it – usually, a byproduct of specific sociopolitical conditions. In other words, it is worth reiterating the platitude that communication may well be hindered notwithstanding language similarity: even speakers of the same variety can have trouble communicating information in a specific register not common to all of them,[1] while, notes Wolff (1959: 35):

> [I]n some areas there is a very low correlation between lexico-structural comparability on the one hand and intelligibility, claimed or proven, on the other. In other words, two dialects might prove to be extremely close when subjected to comparative linguistic analysis, while, at the same time, speakers of these dialects would claim that they could not understand each other.

On the other hand, intelligibility travels across language barriers, so that:

> linguistic (phonemic, morphemic, lexical) similarity between two dialects does not seem to guarantee the possibility of interlingual communication; similarly, the existence of interlingual communication is not necessarily an indication of the linguistic similarity between two such dialects (Wolff 1959: 36).

All of which is very interesting and certainly a nuisance if the task is to measure languageness, but does not preclude it, either logically or empirically.

3.3. "A matter of degree"

In a rather long discussion of intelligibility (many authors are much more dismissive), Hudson (1996: 35), after having mentioned that the criterion of mutual intelligibility "cannot be taken seriously because there are such serious problems in its application", and, repeating the point made in 3.1. above, that "even *popular usage* does not correspond consistently" (emphasis in the original) to it, he goes on:

> Mutual intelligibility is a matter of degree, ranging from total intelligibility down to total unintelligibility. How high up this scale do two varieties need to be in order to count as members of the same language? This is clearly a question which is best avoided, rather than answered, since any answer must be arbitrary.
> […]
> Mutual intelligibility is not really a relation between varieties, but between people, since it is they, and not the varieties, that understand one another. This being so, the degree of mutual intelligibility depends not just on the amount of overlap between the items in the two varieties, but on qualities of the people concerned. One highly relevant quality is *motivation* […] Another relevant quality of the hearer is *experience* (Hudson 1996: 35-36; emphasis in the original)

Intelligibility is certainly a matter of degree (with 100% mutual intelligibility plausibly impossible to reach – if one has to believe Oscar Wilde when he complained being so clever

[1] I thank Ilaria Micheli (University of Trieste) for her suggestions on this point.

that sometimes could not understand himself). And speakers' motivations, past experiences and interests (not to mention sheer linguistic abilities) do exist. But they are empirical issues, to be solved empirically.

Intelligibility may not be a problem for many linguists but it turns out to be a big issue in other fields, ranging from communication technology to medicine; and quite a problem in assessing the accuracy of radio transmission systems and in the definition of hearing impairments.

For a certain sociolinguistic approach to languageness it may be all so sad, but intelligibility *has* been tested and measured, and intelligibility tests have been proposed, discarded, amended; in the end, thresholds have been discussed and agreements have even often been reached.

3.4. The (partially) false problem of dialect chains

> 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 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 not just difficult, but in principle impossible to answer the question how many languages are spoken in the world (Comrie 1987: 2-3).

In short, the problem is:

- if: A & B, B & C, … are mutually intelligible and A & C are not;
- then: (A & B) and (B & C) would be dialects of α;
- this would imply that C is at the same time a dialect of α (as it is intelligible with B) and **not** a dialect of α (as it is not intelligible with A).

Actually, to know how many languages are there in a dialect chain is mathematically easy, as convincingly shown by Hammarström (2008). Without repeating his demonstration, it may suffice here to say that:

> 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 (Hammarström 2008: 4).

This means that, in a group *X* composed of just three members *{A, B, C}*, one can have a single block (i.e., $X = k$): {A, B, C}. This implies that there is mutual intelligibility among all the members of X, which is of course definitionally impossible.

Another theoretical possibility is {A}, {B}, {C}: here, each member of *X* is a block and there is no intelligibility between the members of *X*. Again, this is definitionally impossible.

More interestingly, one of the three following partitions may arise:

1. {A, B}, {C}

2. {A}, {B, C}
3. {A, C}, {B}

Partition 3. is definitionally impossible (we know that A and B are intelligible). This leaves us with two possible partitions.

The number of languages in a chain is therefore unique (here it is two), but there may be several satisfying partitions into *k* blocks; moreover, calculating the total number of *k* blocks increases exponentially with any additional member in the chain.

Summing up:

- if two varieties are the same language, then they are mutually intelligible; but
- if two varieties are mutually intelligible, they are not necessarily varieties of the same language.

Is the problem solved? Not really: we can know the total number of languages, but we do not know the correct partition of their dialects: in our example, is it {A, B}, {C} or {A}, {B, C} correct? Where to put the boundary?

We may know how many languages are there, but not what they are. But, at the very least, we have debunked the myth of dialect continua.

4. Rescuing *Abstand*

There is not much of a *pars construens* in this article: I won't compare, discuss and evaluate different approaches to measuring languageness, and I will restrict myself to mentioning a few recent results and on-going work for what concerns the "contested languages" (Tamburelli & Tosco 2021) of Italy.

Following Gooskens (2007), Tamburelli and Brasca (2017) have recently shown that a dialectometric approach to the varieties traditionally spoken in the northern part of Italy dovetails nicely with traditional subgroupings. Interestingly, the more traditional classifications are marred by purely sociolinguistic analyses – and quite often their accompanying political and ideological underpinnings – the more they are proven wrong when dialectometry is applied. Thus, while the Gallo-Italic grouping in the North of Italy is confirmed, Italo-Romance as an over-arching 'Italian' group is not (Tamburelli & Brasca 2017: 10): as long suspected, Italo-Romance is not a valid genetic grouping – but it can be so in a sociolinguistic sense (: all the languages spoken in a certain area and subject to Italian as a roof language; cf. also Regis 2020). The history of the very concept of Italo-Romance (basically only found in Italian works) exposes its political and ideological biases (unsurprisingly, it is found in De Mauro's 1963 influential *Storia linguistica dell'Italia unita*). Quite to the contrary, Gallo-Italic is revealed to be part of Gallo-Romance, and closer to Occitan than to Italian, while Occitan is actually closer to French than Gallo-Italic is to Italian.

In the meantime, and following Tang and van Heuven (2009) for Chinese 'dialects', Tamburelli (2014) has definitely demonstrated the languageness of, e.g., Lombard by using the SPIN test first proposed by Kalikow et al. (1977). Monolingual Italian speakers with no previous exposure to Lombard were given a selection (18 sentences) of the 'high predictability' sentences of the SPIN test, such as the Lombard translation of *the candle flame melted the wax* or *the workers are digging a ditch*. They were asked to write down the Italian equivalent of the final word only for each stimulus sentence. The results were appalling, with mean intelligibility down at 44.3%, much below the standard threshold for minimal acceptable communication of 75%. Brasca's ongoing work has confirmed Tamburelli's (2014) results, with the intelligibility of the Gallo-Italic speech of the Emilian town of Pavullo down at 38% in the Tuscan town of

Piteglio, which lies to the south across the Appennines but only 68 kms (slightly more than 40 miles) by road, and 34 kms as the crow flies.

Much remains to be done, and many important problems have not even been properly addressed, let alone solved: rampant examples are how to calculate the intelligibility in cases of general mutual bilingualism, and how to deal with the aforementioned (3.2.) question of how to deal with asymmetrical intelligibility: how to measure intelligibility between varieties when speakers are all at least bilingual in a national and related language? This is the case of many minority languages of Italy and other European countries, from Germany to Spain. The answer seems to be that asymmetrical intelligibility is fine: in Pavullo they are likely to perfectly understand the Tuscan variety of Piteglio (very similar to Standard Italian), but the very fact that only 38% of their speech is understood in Piteglio is enough to prove that we are dealing here with separate languages.

Gallo-Italic is not a single language: dovetailing with popular beliefs on differences, Brasca's ongoing research also shows that the intelligibility between single Gallo-Italic varieties falls under the threshold for successful communication, especially when peripheral varieties are considered. Thus, while for speakers of Piedmontese 85% of Lombard is intelligible, for speakers of Lombard the intelligibility of Piedmontese goes down to 70% (Lissander Brasca, p.c., February 21, 2019).

In all these cases (and, we can surmise, countless others across the globe) the lowest figure is all that is needed in order to assess languageness (the highest one has certainly its uses, e.g., in the assessment of bilingualism).

5. Envoi (instead of a conclusion)

Just as measuring the intelligibility between, say, English and Mandarin makes little sense, also a dialectometric approach to these languages will be a colossal waste of time, because zero or a figure close to it is the result. Crucially, dialectometry, as its very name implies, is a tool to measure dialectal difference: it is feasible up to a certain limit, but when whole phonemes (and all the phonemes in a string) are different it becomes impracticable. This does not detract from its usefulness: it is exactly the intricacy of multilingual situations across the globe among a multiplicity of minorities (their 'messiness', for the unfortunate monolinguals of many Western countries who since generations have been the victims of the aggressive linguistic policies of the modern state) that calls for painstaking measurement.

Is this "superdiversity" (Blommaert & Rampton 2012)? Maybe. Certainly, it is the only sensible approach to an assessment of language diversity, which, in its turn, is a prerequisite to salvaging what of it is salvageable (Tosco 2017).

For the time being, we can be content with reiterating that:

- languages do exist. Beyond the veil of political and ideological narratives, languages exist because communication exists; different languages are the result of different and mutually unintelligible solutions to the communication problem.
- languageness is measurable because intelligibility is measurable.
- while *Ausbau*-ization (Tosco 2008) involves the use of linguistic tools with a view to increase the distance of a language (its *Abstand* level) vis-à-vis its neighboring competitors, in the end it is *Abstand* languages that general linguistics deals with.

References

Blommaert, Jan & Rampton, Ben. 2012. "Language and Superdiversity". *MMG Working Paper 12-09* (https://www.mmg.mpg.de/59866/WP_12-09_Concept-Paper_SLD.pdf).

Comrie, Bernard. 1987. “Introduction.” In Comrie, Bernard (ed.), *The World’s major Languages*, 1-22. London: Croom Helm.

De Mauro, Tullio. 1963. *Storia linguistica dell’Italia unita*. Bari: Laterza.

Fasold, Ralph W. 2005. “Making languages.” In Cohen, James & McAlister, Kara T. & Rolstad, Kelly & MacSwan, Jeff (eds.), *Proceedings of the 4th International Symposium on Bilingualism*, 697-702. Somerville, MA: Cascadilla Press.

Gooskens, Charlotte. 2007. “The contribution of linguistic factors to the intelligibility of closely related languages.” *Journal of Multilingual and multicultural development* 28. 445-467.

Hammarström, Harald. 2008. “Counting Languages in Dialect Continua Using the Criterion of Mutual Intelligibility.” *Journal of Quantitative Linguistics* 15(1). 36-45.

Hudson, Richard A. 1996. *Sociolinguistics.* Cambridge: Cambridge University Press.

Kalikow, Daniel N. & Stevens, Kenneth N. & Elliott, Lois L. 1977. “Development of a Test Speech Intelligibility in Noise Using Sentence Materials with Controlled Word Predictability.” *Journal of the Acoustical Society of America* 61. 1337-1351.

Kloss, Heinz. 1967. “‘Abstand languages’ and ‘ausbau languages’”. *Anthropological linguistics* 9(7). 29-41.

Makoni, Sinfree. 2005. “Toward a more inclusive applied linguistics and English language teaching: A symposium.” *TESOL Quarterly* 39(4). 716-719.

Nunberg, Geoffrey. 1997. “Topic... Comments; Double Standards.” *Natural Language and Linguistic Theory* 15. 667-675.

Pennycook, Alastair. 2006. “Postmodernism in Language Policy.” In Ricento, Thomas (ed.), *An Introduction to Language Policy: Theory and Method*, 60-76. Oxford: Blackwell.

Pennycook, Alastair. 2007. “The myth of English as an international language.” In Makoni, Sinfree & Pennycook, Alastair (eds.), *Disinventing and reconstituting languages*, 90-115. Bristol: Multilingual Matters.

Regis, Riccardo. 2020. “Italoromanzo.” *Revue de Linguistique Romane* 84. 5-39.

Tamburelli, Marco. 2014. “Uncovering the ‘hidden’ multilingualism of Europe: an Italian case study.” *Journal of Multilingual and Multicultural Development* 35(3). 252-270.

Tamburelli, Marco & Brasca, Lissander. 2017. “Revisiting the classification of Gallo-Italic: a dialectometric approach.” *Digital Scholarship in the Humanities* 33(2). 442-455.

Tamburelli, Marco & Tosco, Mauro. 2021. “What are contested languages and why should linguists care?” In Tamburelli, Marco & Tosco, Mauro (eds.), *Contested Languages: The Hidden Multilingualism of Europe*, 3-17. Amsterdam: John Benjamins.

Tang, Chaoju & van Heuven, Vincent J. 2009. “Mutual Intelligibility of Chinese Dialects Experimentally Tested.” *Lingua* 119(5). 709-732.

Tosco, Mauro. 2008. “Introduction: Ausbau is everywhere!” *International Journal of the Sociology of Language* 191. 1-16.

Tosco, Mauro. 2017. “On counting languages, diversity-wise.” In Micheli, Ilaria (ed.), *ATrA 3 - Cultural and linguistic transition explored: proceedings of the ATrA closing workshop – Trieste May 25-26, 2016*, 234-245. Trieste: EUT.

Trudgill, Peter. 2000. *Sociolinguistics: An Introduction to Language and Society.* 4th ed. London: Penguin.

Wolff, Hans. 1959. “Intelligibility and inter-ethnic attitudes.” *Anthropological linguistics* 1. 34-41.