# Measurement of linguistic diversity

Shlomo Weber

Center for Study of Diversity and Social Interactions,
New Economic School

BEROC, Minsk, Belorussia
September 28, 2018

# Outline

## Diversity

- General overview
- Measurement of linguistic diversity

## Group Identification

- Group boundaries
- Group association

## Fractionalization indices

- Dichotomous fractionalization indices
- Nondichotomous fractionalization indices
- Linguistic distances

## Polarization indices

## Conclusion

# Outline

## Diversity

### General overview

### Measurement of linguistic diversity

## Group Identification

### Group boundaries

### Group association

## Fractionalization indices

### Dichotomous fractionalization indices

### Nondichotomous fractionalization indices

### Linguistic distances

## Polarization indices

## Conclusion

# There is a wide range of diversity facets:

linguistic
religious
historical
economic
ideological
geographical
genetic
and many others.

# Is diversity good or bad?

> "An angel is more valuable than a stone. It does not follow, however, that two angels are more valuable than the one angel and one stone."
> Thomas Aquinas, *Summa Contra Gentiles, III.*

# Outline

Diversity

- General overview
- Measurement of linguistic diversity

Group Identification

- Group boundaries
- Group association

Fractionalization indices

- Dichotomous fractionalization indices
- Nondichotomous fractionalization indices
- Linguistic distances

Polarization indices

Conclusion

There are two main elements in examining linguistic diversity. One is the set of attributes (native language, age, education, gender, etc.) that exist in the society or the number of languages in of some kind in the nature.
In social sciences we are mostly interested in the sizes of population groups that correspond to various attributes. That is, a distribution of the entire population across the groups population is indispensable for our analysis.

# Outline

Diversity

General overview

Measurement of linguistic diversity

Group Identification

Group boundaries

Group association

Fractionalization indices

Dichotomous fractionalization indices

Nondichotomous fractionalization indices

Linguistic distances

Polarization indices

Conclusion

# Two facets

- Group boundaries – how does one define a partition of the country or countries into separate groups?
- Group association – how do individuals identify themselves with a community to which they belong?

# Group boundaries

A major challenge is the prevalence of multiple identities. People may speak several languages using them in communication across different cultural zones.

To construct an ethnolinguistic map, one can use the dominant linguistic identity. The first attempt of creating a comprehensive world *atlas* was undertaken by Soviet ethnographers in the Miklukho-Maklay Research Institute in Moscow. The result, called ELF (Ethno-Linguistic Fractionalization), was published in Atlas Narodov Mira in 1964. This remarkable dataset was picked by Western scholars, starting with Rustow (1967), Taylor and Hudson (1972) and for almost fifty years played the crucial role in analyzing the impact of linguistic diversity on growth, investment in public goods, quality of government services, corruption, etc.

The identification of distinct languages or dialects may not be straightforward. Are Serbian and Croatian different languages? Should various dialects of Italian, German and Mandarin be treated as separate languages? While economists are not qualified to determine whether Serbian and Croatian are the same language or not, they can mitigate the impact of this determination by using the notion of linguistic proximity of, say, Serbian and Croatian. The notions of linguistic distanc will be discussed later.
Note that if one takes the different dialects of Italian to constitute different groups, then Italy appears to be very diverse. However, if one considers these different dialects to be only minor variations of Italian, then Italy turns to be quite homogeneous.

# Outline

- Diversity
  - General overview
  - Measurement of linguistic diversity
- Group Identification
  - Group boundaries
  - Group association
- Fractionalization indices
  - Dichotomous fractionalization indices
  - Nondichotomous fractionalization indices
  - Linguistic distances
- Polarization indices
- Conclusion

# Group association

While the issue of objective identification is extensively discussed in the economic literature, the question of self-identification requires more attention (Akerlof and Kranton (2000)).

Esteban and Ray (1994) extensively examine the abstract notions of identification with one's own group and alienation towards other groups. In their framework both notions solely depend on the size of the groups.

Aspachs-Bracons et al. (2008) studied the rise of Catalan identity after the introduction of the compulsory bilingual education in 1983.

Castaneda-Dower et al. (2017) conducted an empirical study of the rise of alienation levels of various groups based on the historical patterns of English acquisition in the pre-colonial period in the protracted civil war in Sri Lanka.

But still more empirical research is needed to study individual identity choices to associate them with one linguistic group or another. In the U.S. context, for instance, how strong is the association of individuals with African American Vernacular English (AAVE) and New York Latino English (NYLE)?

Obviously, the identification of individuals is driven by the fear of being rejected by their own community if they choose to speak Standard English instead of the vernacular language that the majority of the community speaks (Lewis, 2007). It could be the case that the US population should be split into three groups: those who learn Standard English as their first language, those who learn a nonstandard dialect of English natively, and those who do not learn English as their mother tongue? (Baugh, 1999).

# Outline

Diversity

General overview

Measurement of linguistic diversity

Group Identification

Group boundaries

Group association

Fractionalization indices

Dichotomous fractionalization indices

Nondichotomous fractionalization indices

Linguistic distances

Polarization indices

Conclusion

# Dichotomous fractionalization indices

The most often used index defined for a multilingual society divided into distinct goups, each member of which speaks the same native language (we disregard the proficiency in other languages).
Let society with the total population of $N$ individuals consist of $K$ groups, $k = 1, \dots, K$.
The population of $k$-th group is given by $N_k$ and $\sum_{k=1}^{K} N_k = N$.
Let $n_k = \frac{N_k}{N}$ be the fraction of the $k$-th group population in the entire society.
We define the index, referred to as $A$-index, as the the probability that two individuals, randomly picked from the entire society, belong to two different group.

Formally, the $A$-index can be presented as

$$A = 1 - \sum_{k=1}^{K} n_k^2.$$

The index was introduced by Gini (1912) as the *mutuality index*. It was later rediscovered by Simpson (1949) and Greenberg (1956), who called it the *monolingual nonweighted index*. It is also a reversed Hirschmann-Herfindahl index often applied for estimating the degree of industrial competiteveness.

Note that in calculating the value of $A$-index we utilize the dichotomous distances. Individuals either belong to same group or do not. In former case their linguistic distance is zero, and in the latter it is one. In doing so, we ignore the challenge of linguistic proximity and simply set 1 for any non-zero linguistic distance.

Another important nondichotomous index is the Shannon (or Shannon-Wiener) (1948) entropy:

$$E = -\sum_{k=1}^{K} n_k \log n_k .$$

The entropy is actually much more often used in biology, statistics and information science, but not in social sciences where the usage of the *A*-index is more prevalent.

Both indices have similar mathematical properties and they were unified through the common axiomatical formulation in Davydov and Weber (2016) (see also Hill (1973) and Simovici and Jaroszewiz (2002)), who offered a general form

$$A^{\alpha} = 1 - \sum_{k=1}^{K} n_k^{\alpha},$$

where $\alpha$ is a positive parameter different from one. Obviously, the value of $A$-index coincides with $A^{\alpha}$ for $\alpha = 2$. It is also quite easy to verify that $A^{\alpha}$ approaches the entropy $E$ when $\alpha$ tends to one.

The value of the parameter $\alpha$ in the above formulation can be interpreted as the degree societal sensitive towards diversity. Societies, regions, cities or counties may differ with the respect to its own value of fractionalization. Some could be threatened by diversity while others may welcome it, thus, exhibiting the attitudes that may have profound economic and politcal outcomes. Ottaviano and Peri (2006) show that Los Angeles, New York and San Francisco have a substantially higher degree of linguistic diversity than, say, midwestern cities Cincinnati and Indianoplis. Much less obvious are differences in "perceived diversity" that indicates how people feel about the impact of globalization, channeled through employment prospects and the presence of immigrants in their own communities. In other words, different societies choose a different $\alpha$, and the identification of that parameter should be an important topic in the research in this field.

# Outline

Diversity

General overview

Measurement of linguistic diversity

Group Identification

Group boundaries

Group association

Fractionalization indices

Dichotomous fractionalization indices

Nondichotomous fractionalization indices

Linguistic distances

Polarization indices

Conclusion

# Nondichotomous fractionalization indices

The reliance on *A*-index may produce unexpected results. Desmet et al. (2009) compared the value of that index in two European countries, Andorra and Belgium.
In a small southeuropean principality of Andiorra, with the population of less than 100,000 peole, roughly a half of its residents have Catalan as the native tongue, whereas the native tongue of the other half is Spanish.
In Belgium the split is about 60 and 40 percent between the Dutch-speaking and the French-speaking populations.

A simple algebra shows that the value of the $A$ index is $1 - 0.5^2 - 0.5^2 = 0.5$ for Andorra and $1 - 0.6^2 - 0.4^2 = 0.48$. In other words, Andorra is more linguistically diverse than Belgium! The reason for this bizarre conclusion that $A$-index does not take into account the proximity between languages. Catalan and Spanish are similar Romance languages, whereas Dutch and French, being members of two distinct language families, Germanic and Romance, are quite distant from each other. The incorporation of linguistic proximity would (and does) make Belgium more linguistic diverse than Andorra.

# Outline

- Diversity
  - General overview
  - Measurement of linguistic diversity
- Group Identification
  - Group boundaries
  - Group association
- Fractionalization indices
  - Dichotomous fractionalization indices
  - Nondichotomous fractionalization indices
  - Linguistic distances
- Polarization indices
- Conclusion

# Language trees

Lexicostatical distances:
The distance matrix is based on cognate date collected by Isidore Dyen at Yale University in the 1960s:
200 basic meanings (chosen by Swadesh (1952))
95 Indo-European speech varieties (languages and dialects)
For each meaning - there is a cognate class of different speech varieties that have an unbroken history of descent from common ancestral word.
For every two varieties, we calculate the number of "cognate" and "non-cognate" meanings. If for example, we have 80 cognate and 120 non-cognate for a pair of languages, the Dyen distance is $\frac{120}{200} = 0.6$.

## Dyen Matrix of distances between the EU25 + RU, UKR languages

| | IT | FR | SP | PT | GE | DU | SW | DA | EN | LI | LA | SV | CZ | SL | PL | GR | RU |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| IT | 0 | 0,20 | 0,21 | 0,23 | 0,73 | 0,74 | 0,74 | 0,74 | 0,75 | 0,76 | 0,78 | 0,76 | 0,75 | 0,75 | 0,76 | 0,82 | 0.76 |
| FR | 0,20 | 0 | 0,27 | 0,29 | 0,76 | 0,76 | 0,76 | 0,76 | 0,76 | 0,78 | 0,79 | 0,78 | 0,77 | 0,76 | 0,78 | 0,84 | 0.77 |
| SP | 0,21 | 0,27 | 0 | 0,13 | 0,75 | 0,74 | 0,75 | 0,75 | 0,76 | 0,77 | 0,79 | 0,77 | 0,76 | 0,76 | 0,77 | 0,83 | 0.77 |
| PT | 0,23 | 0,29 | 0,13 | 0 | 0,75 | 0,75 | 0,74 | 0,75 | 0,76 | 0,78 | 0,80 | 0,78 | 0,76 | 0,76 | 0,77 | 0,83 | 0.77 |
| GE | 0,73 | 0,76 | 0,75 | 0,75 | 0 | 0,16 | 0,30 | 0,29 | 0,42 | 0,78 | 0,80 | 0,73 | 0,74 | 0,74 | 0,75 | 0,81 | 0.76 |
| DU | 0,74 | 0,76 | 0,74 | 0,75 | 0,16 | 0 | 0,31 | 0,34 | 0,39 | 0,79 | 0,80 | 0,75 | 0,76 | 0,75 | 0,77 | 0,81 | 0.78 |
| SW | 0,74 | 0,76 | 0,75 | 0,74 | 0,30 | 0,31 | 0 | 0,13 | 0,41 | 0,78 | 0,79 | 0,75 | 0,75 | 0,74 | 0,76 | 0,82 | 0.75 |
| DA | 0,74 | 0,76 | 0,75 | 0,75 | 0,29 | 0,34 | 0,13 | 0 | 0,41 | 0,78 | 0,80 | 0,73 | 0,75 | 0,73 | 0,75 | 0,82 | 0.74 |
| EN | 0,75 | 0,76 | 0,76 | 0,76 | 0,42 | 0,39 | 0,41 | 0,41 | 0 | 0,78 | 0,80 | 0,75 | 0,76 | 0,75 | 0,76 | 0,84 | 0.76 |
| LI | 0,76 | 0,78 | 0,77 | 0,78 | 0,78 | 0,79 | 0,78 | 0,78 | 0,78 | 0 | 0,39 | 0,66 | 0,62 | 0,60 | 0,64 | 0,83 | 0.62 |
| LA | 0,78 | 0,79 | 0,79 | 0,80 | 0,80 | 0,80 | 0,79 | 0,80 | 0,80 | 0,39 | 0 | 0,68 | 0,67 | 0,64 | 0,67 | 0,85 | 0.64 |
| SV | 0,76 | 0,78 | 0,77 | 0,78 | 0,73 | 0,75 | 0,75 | 0,73 | 0,75 | 0,66 | 0,68 | 0 | 0,34 | 0,31 | 0,37 | 0,82 | 0.39 |
| CZ | 0,75 | 0,77 | 0,76 | 0,76 | 0,74 | 0,76 | 0,75 | 0,75 | 0,76 | 0,62 | 0,67 | 0,34 | 0 | 0,09 | 0,23 | 0,84 | 0.26 |
| SL | 0,75 | 0,76 | 0,75 | 0,76 | 0,74 | 0,75 | 0,74 | 0,73 | 0,75 | 0,60 | 0,64 | 0,31 | 0,09 | 0 | 0,22 | 0,83 | 0.26 |
| PL | 0,76 | 0,78 | 0,77 | 0,78 | 0,75 | 0,77 | 0,76 | 0,75 | 0,76 | 0,64 | 0,67 | 0,37 | 0,23 | 0,22 | 0 | 0,84 | 0.27 |
| GR | 0,82 | 0,84 | 0,83 | 0,83 | 0,81 | 0,81 | 0,81 | 0,82 | 0,84 | 0,83 | 0,85 | 0,82 | 0,84 | 0,83 | 0,84 | 0 | 0.83 |
| RU | 0,76 | 0,77 | 0,77 | 0,77 | 0,76 | 0,78 | 0,75 | 0,74 | 0,76 | 0,62 | 0,64 | 0,39 | 0,26 | 0,26 | 0,27 | 0,83 | 0 |
| UKR | 0,77 | 0,78 | 0,78 | 0,78 | 0,76 | 0,79 | 0,76 | 0,76 | 0,78 | 0,63 | 0,64 | 0,36 | 0,24 | 0,19 | 0,20 | 0,77 | 0,22 |
| BLR | 0,76 | 0,78 | 0,77 | 0,77 | 0,75 | 0,76 | 0,75 | 0,74 | 0,76 | 0,63 | 0,63 | 0,39 | 0,29 | 0,22 | 0,25 | 0,82 | 0,27 |

IT - Italian; FR - French; SP - Spanish; PT - Portugal; GE - German; DU - Dutch; SW - Swedish; DA - Danish; EN - English; LI - Lithuanian; LA - Latvian; SV - Slovenian; CZ - Czech; SL - Slovak, PL - Polish; GR - Greek.

Note als that the distance between the Belorussian and Ukrainian is 0,16.

Greenberg (1956) indroduced a monolingual weighted index, which, in simple words, accounts for an average linguistic distance between randomly chosen individuals within the society, In addition to the notation of the previous subsection, let $d_{ki}$ denote the linguistic distance bertween two gtoups $i, k, = 1. \ldots, K$. The distance could be derived via any of the methods described earlier in thuis section. Then we have index $B$ whose formal presentation is given by

$$B = \sum_{k=1}^{K} \sum_{i=1}^{K} n_k n_i d_{ki}.$$

It is quite easy to see that $B$-index is a generalization of the $a$-index, dischotomous distances are replaced by an arbitrary distnace metric. Indeed, if we impose a dichotomy on $A$, i.e., assume that $d_{ki} = 0$ if $i = k$ and $d_{ki} = 1$ if $i \neq k$, then $B$ turns into $A$:

$$B = \sum_{k=1}^{K} n_k (1 - n_k) = 1 - \sum_{k=1}^{K} n_k^2.$$

It worth to pointing out that the emprical analysis relying on the $B$ index requires a more extensive dataset than simply using the index $A$. However, the effort could be worth it, as Desmet et al. (2009) in their cross-country analysis of redistribution patterns, clearly indicate a much stronger explanatory power of index $B$. The importance of incorporation of linguistic distances in the definition of societal indices is also supported by Dower et al. (2017) in the context of their analysis of the Sri Lanka conflict.

Of particular importnace is center-periphery relations or even tension between the centeral and prerepheral regions. To highlight this point, Desmet et al. (2009, 2017) have assigned a special role to one of the regions, say, region 1, called the "center" and denoted by $c$. The other $K-1$ regions are assumed to be "peripheral". In calculating a variant of $B$-index, only the distances between the center and perphery are accounted for. whereas the bilateral links between any two perpheral regions are disregarded. That is,

$$CP = n_c \sum_{k=2}^{K} n_k d_{kc}.$$

# Polarization indices

The notion of polarization, and the indices it generates, adds an additional element to the build-up that led to the introduction of *A*- and *B*-index indices. To recall, both of those indices rely on the notion of pre-existing partition into distinct linguistic groups. The polarization approach adds an important facet of individual self-identification for members of the society. The self-identification comes through in two ways. One, is the strength of identification with others in one's own group, another is alienation toward the others.

Esteban and Ray (1994) defined a notion of *social effective antagonism* that combines both identification and alienation, that depend only on the size of the groups. Esteban and Ray examine income pollarization when the groups are identified by their income levels and the distances are income differentials between the groups. The fuctional form of their index bult on axiomatic foundations, is close to that of the index $B$:

$$P = \sum_{k=1}^{K} \sum_{i=1}^{K} n_k^{1+\alpha} n_i d_{ki},$$

where $\alpha$ is a positive parameter ranging between 1 and 1.6.
For $\alpha = 0$, the index $P$ is a Gini coefficient of income inequality.

Consequently, Geng (2012) has imposed additional axioms to sprink the range of $\alpha$'s to a single point, $\alpha = 1$.
The Esteban and Ray approach can be adjusted to incorporate linguistic distanced (Montalvo and Reynal-Querol (2005), Desmet et al. (2017) and Dower et al. (2017)). For $\alpha = 1$, Reynal-Querol (2002) offered a dichotomous version of this index that assumes that $d_{ki}$ is equal to zero if $k = i$, and $d_{ki}$ is equal to one $k \neq i$. The Reynal-Querol index then obtains the following functional form

$$RQ = \sum_{k=1}^{K} \sum_{i=1}^{K} n_k^2 n_i = \sum_{k=1}^{K} n_k^2 (1 - n_k).$$

It is worth pointing out the intuitive difference between two dichotomous indices $A$ and $RQ$. To recall, index $A$ is determined by the probability that two randomly chosen individuals belong to two different groups. Thus, the value of $A$ is the sum of the terms $n_k(1 - n_k)$, each identifying the probabily that one individual belong to group $k$, while the other does not. The value of $RQ$ is determined by the probability that among three randomly chosen individuals two belong to one group, while the third belongs to other. Thus, $RQ$ is represented by the sum of the terms $n_k^2(1 - n_k)$, that is, two individuals belong to group $k$, while the third does not.

Moreover, in the Esteban and Ray model the identification and alienation depend only on the sizes of relevant groups. One may assume that some additional factors, such as linguistic proximity or historical path could play an important part in determining the degree of identification and alienation. In their sttudy of the protracted war in Sri Lanka, Dower et al. (2017) introduce an ethnolinguistic polarization measure that takes into account the impact of historical factors on inter-group relations driven by different patterns of English language acquisition in the colonial era. They used the Reynal-Querol variant of the Esteban-Ray index (both in dichotomous and nondichotomous forms) by comparing its value across all dostricts $j$:

$$D^j = \sum_{k=1}^{K} \sum_{i=1}^{K} (n_k^j)^2 n_i^j d_{ki},$$

where $n_k^j$ and $n_k^j$ denote the fraction of linguistic groups $k$ and $i$, respectively in district $j$, whereas $d_{ki}$ is the linguistic distance between groups $k$ and $i$.

However. that linguistic takes into account changes groups' English proficiency in the precolonial period. By examining the protracted war in Sri Lanka and applying that measure to a dataset on victims of terrorist attacks by district and war period, they find that increasing the share of English speakers resulting from colonial times in each district would result in increasing the number of war victims.

# Conclusion

It is important to point out:

• Fearon and Laitin (2003), Collier and Hoeffler (2004) indicate that $A$-index does not provide an obvious link for likelihood of civil conflicts. The analyssis here requires an alternative index. and, indeed, Montalvo and Reynald-Querol (2005) argue that the polarization index $RQ$ does explain the incidene of civil wars. Moreover, Montalvo and Reynald-Querol (2002) show that a higher level of $RQ$ points out to a longer civil conflicts.

• The fractionalization index $A^{\alpha}$, defined earlier in this section, may indicate that different societies exhibit different levels of $\alpha$. There are different attitudes towards immirgants and the importance of diversity for well-being varies across the communities. It is up to a researcher to identify a proper value of $\alpha$ for each community, while refuting the approach that considers all societies being equal.

- Desmet et al. (2015) construct various linguistic partitions for examining various question. They argue that the data should determine which level of aggregation to select and show that with regard to civil conflict and redistribution, deeper cleavages, and, thus, coarser partitions are more significant. That is, the historical path of language development matters. In contrast, for economic growth and provision public goods, the diversity measure based on more disaggregated classifications of linguistic groups, capturing finer distinctions between languages, are important correlates of growth and public goods provision both in terms of statistical significance and in terms of economic magnitude.