ANALYZING THE LOCALIZATION OF LANGUAGE FEATURES WITH COMPLEX SYSTEMS TOOLS AND PREDICTING LANGUAGE VITALITY

Samuel F. Omlin

Abstract: *Half of the world's languages are in danger of disappearing before the century ends. Efficient protection of these languages is difficult as their fate depends on multiple factors. The role played by the geographic situation of a language in its survival is still unclear. The following quantitative study focused on the relation between the 'vitality' of a minority language and the linguistic structure of the neighboring languages. A large sample of languages in Eurasia was considered. The languages were described based on a complex system of typological features. The spatial distribution of the language features in the sample area was measured by quantifying deviations from purely random configurations. Interactions between the linguistic features were revealed. The obtained interaction network permitted to define a location "quality" index for a language localization. This index was put in relation to corresponding vitality estimations from Unesco. A significant relation could be established between these two variables. The degree of endangerment of the minority languages studied seems effectively related to the linguistic structure of their neighboring languages. Beyond the particular context of endangered languages, the proposed approach constitutes a promising tool to gain more knowledge about the mechanisms that control the geographical distribution of linguistic features.*

Keywords: *Language competition, Complex systems, Interactions, Spatial distribution, Typological language features.*

ACM Classification Keywords: *I.m Miscellaneous; J.5 Arts and Humanities – Linguistics; H.2.8 Database Applications – Data mining, Scientific databases, Spatial databases and GIS.*

Introduction

Numerous languages of the world are endangered: Unesco [2010] estimates that half of the about 6700 languages spoken today will have disappeared before the century ends, if no significant measures are taken. However, efficient revitalization of endangered languages is a challenging task as the fate of a language depends on a myriad complex factors.

In the business world, numerous factors determine the success or failure of an enterprise. Yet, a well-known business doctrine states that the location alone determines, in most cases, whether or not a store may survive the

competition. This study attempts to determine if this doctrine can be applied to the survival of languages, i.e. if the localization of a language is a determinant factor in predicting the success of a language in competition with its neighboring languages. In fact, the following brief overview of recent models of language competition and some comparison to linguistic literature show that the role played by the geographic situation of a language in its ultimate survival, and in particular the role of the linguistic structure of the languages neighboring it, is still unclear.

Abrams and Strogatz [2003] proposed a simple model that describes how two neighboring languages are "competing for speakers". Their article received the attention of numerous researchers because "its fit to the empirical data was exceptional" [Wang and Minett, 2005, p. 265]. The proposed model predicted the death of the weaker language in every case. Many other language competition models followed. While some produce similar results [e.g. Castelló, Eguíluz, and Miguel, 2006, Castelló et al., 2007], others lead to the conclusion that under certain circumstances, minority languages can live in stable coexistence with stronger ones. Models allowing stable coexistence notably focus on historical or geographical factors [Patriarca and Leppänen, 2004, Patriarca and Heinsalu, 2009], include the pride of speakers to their linguistic identity [Schulze and Stauffer, 2006], or take into account linguistic similarity between two or more languages "in competition" [Mira and Paredes, 2005, Teşileanu and Meyer-Ortmanns, 2006]. The following comparison of these results with linguistic literature focuses on the role of the linguistic structure of languages in competition.

Mira and Paredes [2005, p. 1] who studied the coexistence of Galician (in northern Spain) with the dominant Castilian, concluded that two languages can coexist if they are "similar enough". Teşileanu and Meyer-Ortmanns [2006] drew the same conclusion about two or even three neighboring languages. These results are to some extent in agreement with Wardhaugh [1987, p. 17], who notes that in certain situations "there may be little pressure from one language on the other or others". Other linguists affirm that a high linguistic similarity between a minority language and a stronger one can "retard" its assimilation, notably based on the case of the Galician [Monteagudo and Santamarina, 1993], as well as on an a example in the Netherlands [Palmer, 1997] and in Italy [Posner and Rogers, 1993]. However, Posner and Rogers [1993, p. 55] consider this case as an exception and state: "The greater the linguistic distance between languages the less likely is language shift to occur" [similarly Mackey, 2001].

A group of experts from Unesco estimated the degree of endangerment of the living languages in the *Atlas of the World's Languages in Danger* [Moseley, 2009], referred to as *language vitality*, with nine criteria. However, in none of them was geography directly implied. This stresses the importance of clarifying the influence of geographic factors in the context of language endangerment.

The following quantitative study focused on the relation between the vitality of a minority language and the linguistic structure of the languages neighboring it. For this purpose, a mathematical method, having its origins in the economical sciences and identifying optimal localizations to implement commercial stores with empirical success [Jensen, 2006, 2009], was adapted. In fact, Jensen had developed an approach to study the distribution of commercial activities in a city – a network on a heterogeneous geographic space. Similarly, the world is home to a network of languages, or more precisely, of linguistic features. The adapted approach was applied to a big sample of Eurasian languages, the linguistic structure of which was quantified based on a complex system of typological features. The present study is the first, known to the author, to integrate realistic linguistic features in order to describe languages in competition. Teşileanu and Meyer-Ortmanns [2006] had stressed the importance of such an approach. It is also the first to consider a large-sized language network in this context.

After this introduction, the paper is structured as follows: the next section presents briefly the studied language sample and its modeling; section 3 summarizes Jensen's approach [2006, 2009]; section 4 explains an approach to measure the spatial distribution of linguistic features; section 5 shows how the vitality of minority languages can be predicted based on these measurements; section 6 presents the most important results; the last section concludes the study and lists future work to do.

2 Sample and Modeling

A sample of 105 living languages in Eurasia was considered. For this study, the definition and usage of the term *language* of Ethnologue, 16th edition [Lewis, 2009] was used. This allowed referring to a language by a unique key, the code ISO 639-3. The *World Language Mapping System*[1] [Global Mapping International and SIL International, 2010; abbreviated as *WLMS]* permitted to describe the geographical area where the considered languages are spoken and to separate these languages into 186 linguistic communities with *independent vitality*. The term *linguistic community* was defined as the ensemble of speakers of a language in the same country. In fact, speakers of the same language but living in different countries do not have the same political and social environment, such that the speakers of a country can be considered as an independent linguistic community with an independent fate. Such a linguistic community could be identified by the unique code ID-ISO-A2 available in WLMS. This key is a concatenation of the code ISO 639-3 to identify a language and the country code ISO A2. The linguistic structure of a language was quantified based on the database *Jaziky mira* [Academia and Indrik, 1993-2010; abbreviated as *JM*], providing a complex hierarchical system of typological features. Every

[1] Version 16 (elaborated based on Ethnologue, 16th edition)

considered language was characterized by a certain number of *binary typological features*[1]. For the considered linguistic communities, the description of their language in JM could be univocally attributed via the code ISO 639-3. Additionally, a number of speakers (provided by WLMS) was associated and where possible as well a *vitality grade*[2] from the *Atlas of the World's Languages in Danger* [Moseley, 2009; further referred to as *Unesco's atlas*]. The attribution of vitality grades to linguistic communities was a complex and fastidious task as Unesco does not use a standardized key in order to refer to them. In addition, it eventuated that about two thirds of the examined *central points*[3] from Unesco could not be linked to an area from WLMS (nor to a set of areas). The first step consisted in extracting the 186 central points from Unesco's atlas that refer to one single language (i.e. having associated one and only one code ISO 639-3), which is described in JM (the other languages were not of interest for this study). Then, for 43 linguistic communities a clear link to a vitality grade could be established on the basis of some geographic criteria (mainly: global geographic situation, distance of the Unesco's central point to the area and central point of WLMS) as well as on a comparison of the metadata of the two databases WLMS and Unesco's atlas (name(s) of the attributed language, number of speakers, textual description of localization, etc.). 31 of these communities were in the chosen *sample region* (explained in the following). The modeling described above is summarized in the (spatial) UML schema in figure 1.

The three most important objectives when defining the sample were the following: the sample has to be in a *continuous geographic region*; it should represent an *inventory* of linguistic communities of the chosen region *as complete as possible*, especially of the communities with many speakers (evidently, a community could only be included in the sample if it could be described by JM); nevertheless, the sample should be of a *considerable size*, in order to a priori allow gaining statistically significant results. As the two first objectives were concurrent to the last one, an optimal compromise was aimed at. The retained sample region consists of nearly entire Europe, a major part of Asian Russia and a few adjacent regions, in particular the region of the Caucasus Mountains – a "hotspot" of endangered languages – together with its close surroundings.

---

[1] An example of such a binary topological feature is the following: "The word order in the simple phrase is subject, followed by verb and then by object." For this feature, English obtains the value '1' or 'True', Turkish the value '0' or 'False'. This example does not exist exactly like this in JM. It was invented for allowing an easy explanation of the structure of this database. Only the leaves of the hierarchical tree of JM were considered for characterizing the structure of the considered languages.

[2] Unesco attributed to every inventoried linguistic community one of the following ordinal vitality grades: "extinct", "critically endangered", "severely endangered", "definitely endangered" or "vulnerable". When a language is constituted of several independent communities, an *overall vitality* was attributed to the language.

[3] Represents the centre of the area where the speakers live or the coordinates of the largest city or village.

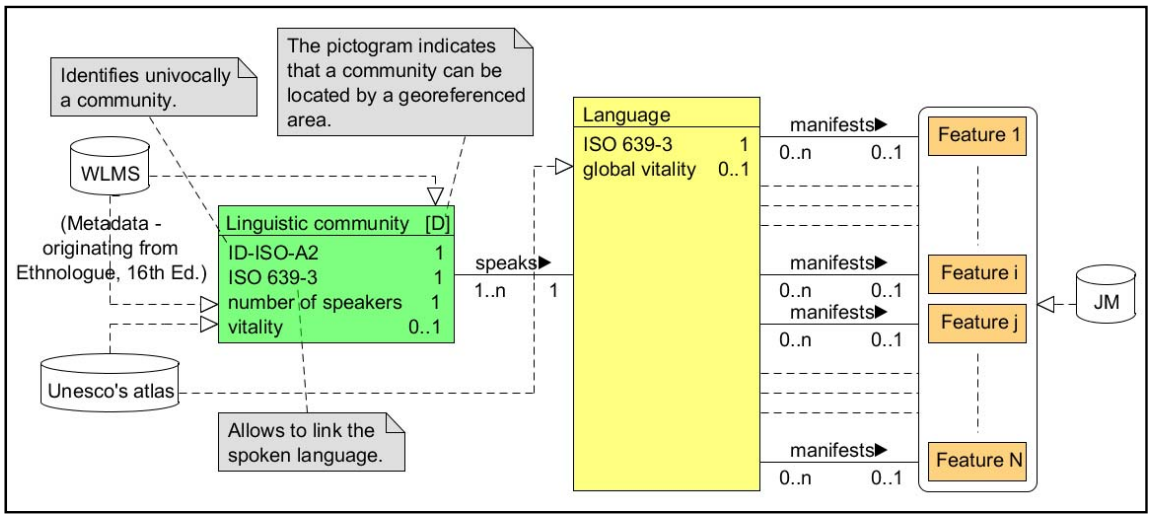

Figure 1: UML schema summarizing the modeling.

3 Analyzing a network on a heterogeneous geographic space

Jensen [2006, 2009] developed an approach to measure the spatial organization of commercial activities in a city. Concretely, he conceived an *M index* to quantify the geographic aggregation and dispersion tendencies of categories of stores. Jensen [2006, p. 1] explained:

The definition of M_{AB} at a given distance r is straightforward: draw a disk of radius r around each store (s) of category A, count the total number of stores (n(s)), the number of B stores ($n_B(s)$), and compare the ratio $n_B(s)/n(s)$ to the average ratio N_B/N, where N refers to the total number of stores in town. If this ratio, averaged over all A stores, is larger than 1, this means that A "attracts" B, otherwise there is repulsion between these two activities. I have chosen r = 100m as this represents a typical distance a customer accepts to walk to visit different stores.[1]

[1] In this quotation Jensen's notation of the variables has been slightly modified (the variable s has been inserted for referring to a store and "*n(S)*", "$n_B(S)$" and "***N***" have replaced "n_{tot}", "n_B" and "N_{tot}" of Jensen's notation). This was done to allow using the presented notation in the explicit formulas that follow.

The ratio $n_B(s)/n(s)$ can be understood as the *concentration* of activity B in the *neighborhood* of A and N_B/N as the *overall concentration* of activity B in the entire city [Prof. Pablo Jensen, oral communication, 2010].

After having presented the basic idea of the M index, Jensen [2009] presented two explicit formulas for the M index, *M intra* and *M inter* (M_{AA} and M_{AB}):

$$M_{AA} \equiv \frac{\frac{1}{N_A}\sum_{s\in S_A}\frac{n_A(s)}{n(s)}}{\frac{N_A-1}{N-1}} \tag{1}$$

$$M_{AB} \equiv \frac{\frac{1}{N_A}\sum_{s\in S_A}\frac{n_B(s)}{n(s)-n_A(s)}}{\frac{N_B}{N-N_A}} \tag{2}$$

where S_A refers to all A stores in the entire city, $n_A(s)$ represents the number of A stores in the neighborhood of store s and N_A stands for the number of A stores in the entire city. The elements in the formula for M inter that were not explained above constitute some corrections that only are of importance in extreme cases. The basic idea of the index remains the same. The basic idea behind M intra is analog to the one of M inter. It can also be interpreted similarly: "if the observed value of the intra coefficient is greater than 1, we may deduce that A stores tend to aggregate, whereas lower values indicate a dispersion tendency" [Jensen, 2009, p. 13]. The interpretations of M intra and M inter are based on the fact that under *pure randomness hypothesis* $E[M_{AA}]$ and $E[M_{AB}]$ equal 1 for all $r>0$.

Based on this M index, Jensen defined a location "quality" index ($Q_A(x,y)$) for a commercial activity A at a point (x,y) as:

$$Q_A(x,y) \equiv \sum_{B\in Act} a_{AB} n_B(x,y) \tag{3}$$

where Act is the set of all considered commercial activities, $n_B(x,y)$ represents the number of neighbor stores of category B around (x,y), $a_{AB}\equiv\log(M_{AB})$ for $A\neq B$ and $a_{AB}\equiv\log(M_{AA})$ for $A=B$. Jensen [2009, p. 18]

explained on the example of bakeries: "The basic idea is that a location that gathers many activities that are 'friends' (i.e. that attract bakeries) and few 'enemies', might well be a good location for a new bakery." The *location quality* for an existing store can be computed by removing it from town and calculating *Q* at its location [Jensen, 2006, 2009].

4 Measuring the spatial distribution of linguistic features

Jensen's M index [2009] was adapted in order to quantify tendencies of typological language features to aggregate or disperse. The *neighborhood of a linguistic community* was defined as the set of communities overlapping its area enlarged by a buffer of size *r* (the area of a community can be contained entirely by another one…). For easiest understanding the main idea of the index is explained in Jensen's words: for each community (*c*) manifesting the feature *A*, sum the number of speakers of all communities in its neighborhood ($n(c)$), sum the number of speakers of the communities manifesting the feature *B* in its neighborhood ($n_B(c)$), and compare the concentration $n_B(c)/n(c)$ to the overall concentration N_B/N, where *N* refers to the sum of the number of speakers of all communities in the entire sample region. If this ratio, averaged over all communities manifesting feature *A* (where the average is weighted by their number of speakers), is larger than 1, this means that *A* "attracts" *B*, otherwise there is repulsion between these two features. The buffer size *r* has been chosen 1 degree (≈110 km). This size lies somewhere in between the maximal commute distance to work undertaken on daily basis and the distance that can be reached by car, train or autobus on a day trip. This choice may appear somewhat arbitrary, but the model seems quite robust to the variations of this parameter [Prof. Jensen, oral communication, 2010].

The spatial distribution of the language features was measured by counting speakers manifesting certain features rather then simply counting communities manifesting them. In fact, this allowed measuring the distribution with higher precision. It seemed important since the number of speakers of the considered communities vary very strongly.

The explicit formulas M intra and M inter were adapted:

$$M_{AA} \equiv \frac{1}{N_A} \sum_{c \in C_A} n_{self}(c) \frac{\frac{n_A(c)}{n(c)}}{\frac{N_A - n_{self}(c)}{N - n_{self}(c)}} \tag{4}$$

$$M_{AB} \equiv \frac{\frac{1}{N_A} \sum_{c \in C_A} n_{self}(c) \frac{n_B(c)}{n(c) - n_A(c)}}{\frac{N_B}{N - N_A}} \tag{5}$$

where C_A refers to all communities manifesting feature A in the entire sample, $n_A(c)$ represents the sum of the number of speakers of all communities manifesting feature A in the neighborhood of the community c, $n_{self}(c)$ equates to the number of speakers of community c itself and N_A represents the sum of the number of speakers of all communities manifesting feature A in the entire sample region ($N_A \equiv \sum_{c \in C_A} n_{self}(c)$). The elements in the formula for M_{AB} that were not explained above constitute the analogue corrections to the ones Jensen (2009) made. The differences of the adapted formula for M_{AA} to Jensen's original one were necessary to achieve the analogue corrections[1]. Also in these adapted formulas the corrections are only of importance in extreme cases and the basic idea of the index remains the same. Likewise, the main idea behind M intra is analog to the one of M inter and can again be interpreted similarly: if the observed value is superior to 1, it means that communities manifesting the feature A tend to aggregate, whereas inferior values indicate a dispersion tendency. As for Jensen's formulas, under pure randomness hypothesis, $E[M_{AA}]$ and $E[M_{AB}]$ equal 1 for all $r > 0$. This justifies the given interpretations of the adapted M index.

Jensen [2006, 2009] interpreted commercial activities that tend to aggregate as "friends", and such that tend to disperse as "enemies". In fact, Jensen [2009] argued that most existing stores are located at places that are "friendly" to them because badly situated stores would perish quite fast. In other words, the spatial distribution of the commercial activities seems to unravel interactions that favor or disfavor *successful* local coexistence of certain activities. From the spatial distribution of language features only, however, it cannot be directly determined which features *successfully* coexist and which do not, as at least half of the languages of the world are endangered and are therefore unlikely to be located in places that are friendly to them. Nevertheless, it seems

[1] Jensen [2009] explained that when measuring the local concentration of activity ***A*** around a store ***s*** of category *A* ($n_A(S)/n(S)$), it has to be compared to a reference concentration that does not take into account this particular store ***s***. This reference concentration is obtained by subtracting 1 (for the store ***s***) from the numerator and the denominator of the overall concentration of activity ***A***. Thus, for every ***A*** store the reference concentration is $(N_A-1)/(N-1)$. In consequence, when averaging the ratio between the local concentration and the reference concentration over all ***A*** stores in the city, the latter concentration appears as a constant and can be factored out. When measuring the local concentration of a feature ***A*** in the neighborhood of a community c manifesting this feature *A* ($n_A(C)/n(C)$), in analogy, it has to be compared to a reference concentration that does not take into account this particular community ***c***. This reference concentration is obtained by excluding the particular community c from consideration. In other words, this community's number of speakers ($n_{self}(C)$) has to be subtracted from the numerator and the denominator of the overall concentration of feature ***A***. Therefore, for every feature ***A*** the reference concentration is $(N_A-n_{self}(C))/(N-n_{self}(C))$. In consequence, when averaging the ratio between the local concentration and the reference concentration over all communities c manifesting feature A in the entire sample area, the latter concentration does not appear as a constant like in Jensen's case (as $n_{self}(C)$ is variable), i.e. it cannot be factored out.

that in Eurasia, one can quantify interactions favoring or disfavoring *successful* coexistence between features by considering only communities that are probably not endangered when computing the M index, as they represent most of the speakers of the region. To this modification of the M index is referred to with *C index* in the following. More precisely the C index was defined as follows:

$$C_{AB} \equiv \begin{cases} M_{AA} & \text{for } A = B \\ M_{AB} & \text{for } A \neq B \end{cases} \quad (6)$$

where the considered linguistic communities are only the ones that are probably not endangered. The next section shows that quantifying this *coexistence ability* between features could be as much of interest as measuring their effective spatial aggregation or dispersion.

5 Predicting the vitality of minority languages

A location quality index for a feature, similar to the one Jensen [2009] had conceived for commercial activities, was defined:

$$Q_A(area) \equiv \frac{1}{\sum_{B \in F} n_B(area)} \sum_{B \in F} a_{AB} n_B(area) \quad (7)$$

where F is the set of all considered language features, $n_B(area)$ represents the sum of the number of speakers of the communities manifesting the feature B in the neighborhood of the given $area$ and a_{AB} is defined as follows:

$$a_{AB} \equiv \begin{cases} C_{AB} - 1 & \text{for } C_{AB} \geq 1 \\ -\left(\frac{1}{C_{AB}} - 1 \right) & \text{for } C_{AB} < 1 \end{cases} \quad (8)$$

The adapted Q_A index simply represents the average ability of a feature A, manifested by a community located on the given $area$, to coexist with the features of its neighboring communities (The transformation of C_{AB} allows to have disabilities to coexist – C_{AB} values inferior to 1 – on a same scale as positive abilities – C_{AB} superior to

1.). The location quality for a feature manifested by a community can be computed by removing the community from the sample and calculating Q_A for the *area* associated with this community.

The index is independent from the number of speakers in the neighborhood, i.e. independent from the number of network-entities around the *area*. This is the main difference to Jensen's index, which is amplified by the density of commercial activity in the neighborhood of *(x,y)*, i.e. amplified by number of network-entities around *(x,y)* (this can be verified mathematically by multiplying $n_B(x,y)$ of every point *(x,y)* by a same constant $k \in]0, \infty] \mid k \neq 1$). The reason for this difference is that for a linguistic minority community, a high population density in its neighborhood is susceptible to have a *normally rather negative* influence on its vitality: linguistic minority communities isolated on islands are in general less endangered than the ones situated on the continent [Sutherland, 2003]. As the influence of population density on vitality is not clear, the index was defined completely independent of this factor.

Based on this index the question whether and to what extent the linguistic structure of the communities in the neighborhood of a minority community allows predicting its level of endangerment can be illuminated: for each community its vitality can be compared to the localization qualities of the features it manifests (where a vitality can be assigned, of course). Due to the fact that the 31 considered communities share only few features and therefore also few predictors, for each community, the localization qualities of its features were aggregated (by averaging the z-scores) to form one single location quality index. These 31 community location quality indices were then put in relation to Unesco's corresponding vitality grades.

6 Results

The proposed M and C indices lead to results that are in agreement with visual verifications done. To give an example, figure 2 (next page) shows the spatial distribution of the communities manifesting the features having the ids 3686[1] and 760[2] in JM in the sample region (these features are not manifested in the omitted eastern part of the sample region). On this map, a strong spatial repulsion can be observed between these features. The computed M inter value is about 0.001, which means that in the neighborhood of the speakers manifesting the feature 3686, the average concentration of the speakers using the feature 760 is about a thousandth of the

[1] In JM described as "2.5.3.SIMPLE SENTENCE -> marginal constructions -> Affective"

[2] In JM described as "2.1.4.SYLLABLE -> the element following the vowel -> not more than one consonant"

overall concentration in the entire sample area. The visually observed repulsion seems coherently expressed by the computed M inter value.

Spearman's rang correlation between the (ordinal) vitality grades and the computed location qualities of the 31 considered minority communities has the value of 0.62 (p-value: 0.00009). In consequence, it seems that for the studied sample, the vitality of a minority language is indeed related to the linguistic structure of the neighboring languages. Besides, for all identified endangered communities, with the exception of five, the computed location quality Q_A is below zero, i.e. inferior to the one of an average community. On the other hand, three quarters of the communities that were judged to be probably not endangered (48), obtained a Q_A score above zero, i.e. superior to the one of an average community. It has to be noted though, that there is a certain circularity in the computation of the location quality indices for these latter communities as they had constituted the reference for the computations of the C_{AB} index on which Q_A is based. Figure 3 (next page) shows the score of the location quality computed for the above mentioned linguistic communities in function of their vitality. The 31 considered minority communities are represented as blue circles, the 48 communities that are probably not endangered as red squares.

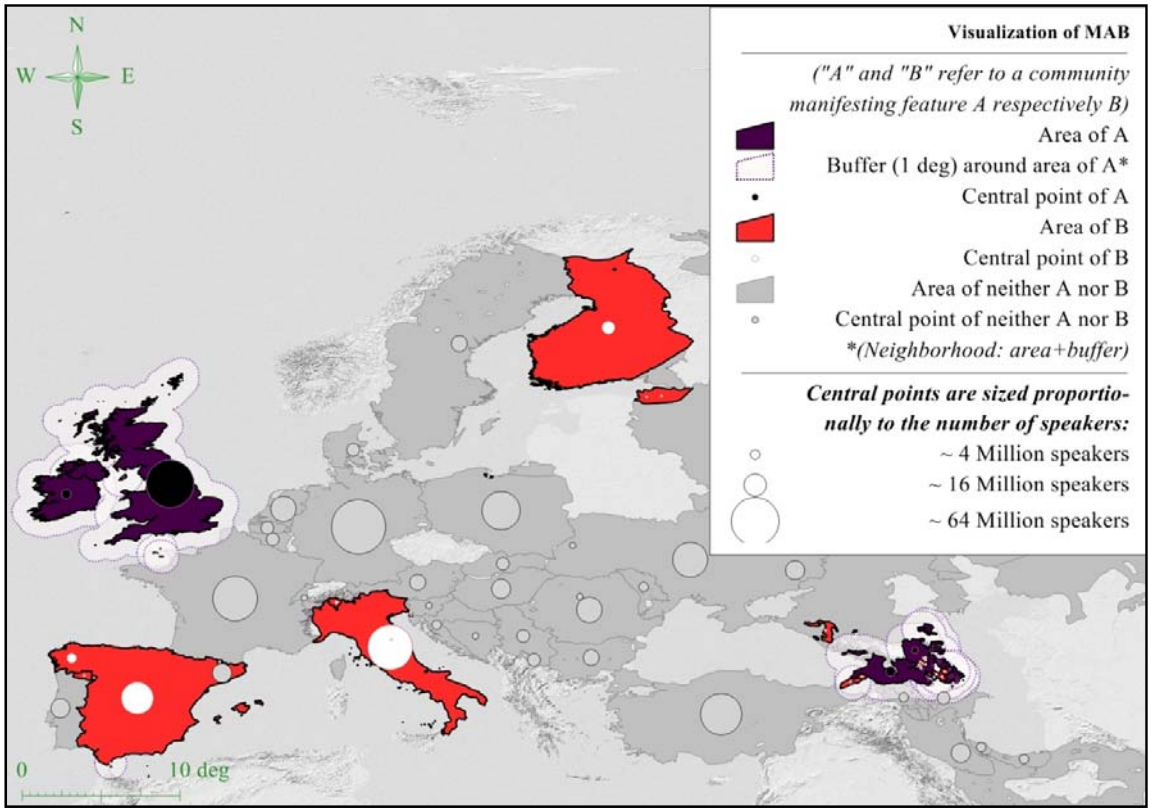

Figure 2: Visualization of M_{AB} computation for the features 3686 and 760.

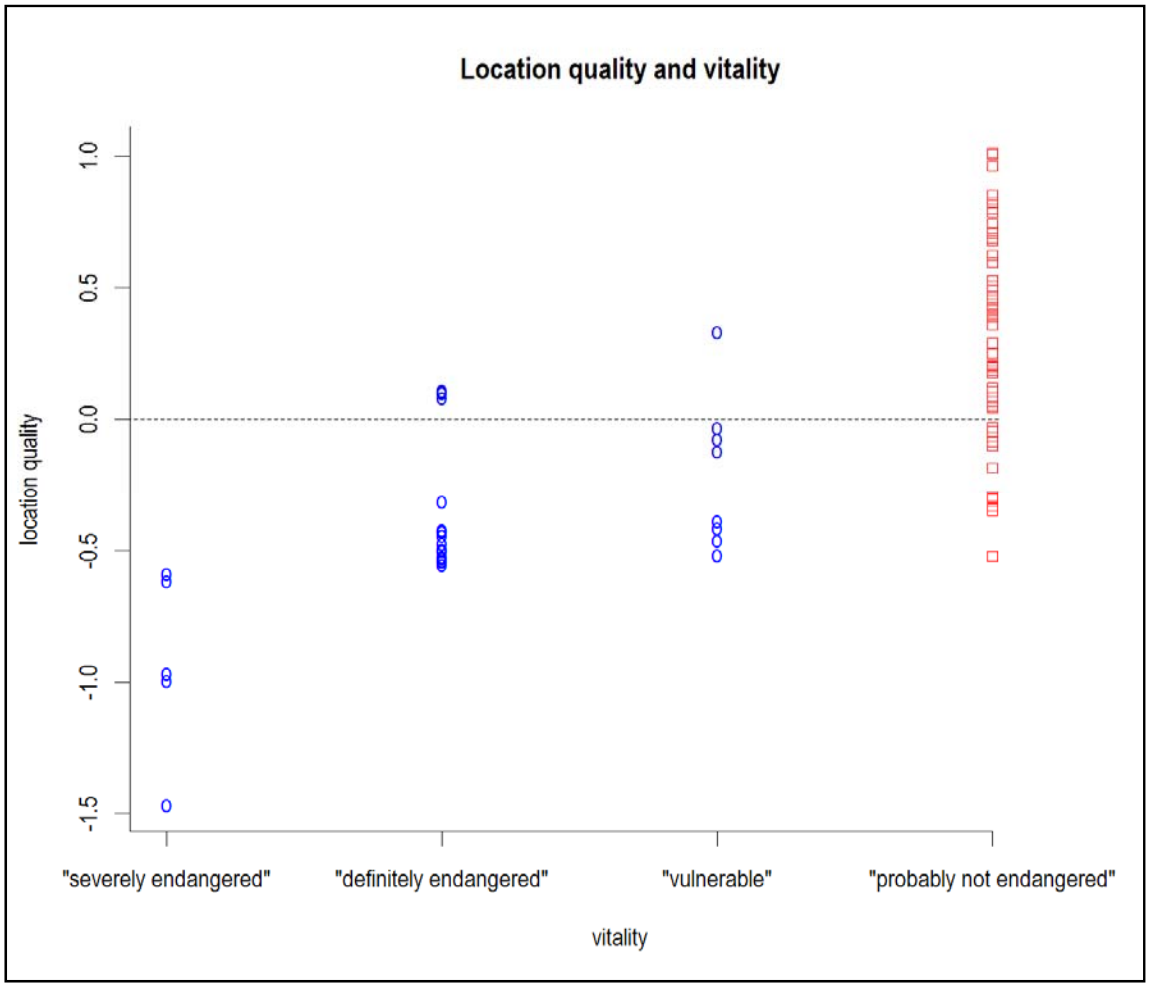

Figure 3: Location quality and vitality of linguistic communities.

Conclusions and future work

Studying the literature on the vast subject of endangered languages has allowed pointing out the importance of researching the influence of the geographic situation of a language on its survival and in particular the role and the importance of linguistic structures in this context. The main result obtained in this study confirms the relevance of the research questions raised: the degree of endangerment of the considered minority languages seems effectively related to the linguistic structure of their neighboring languages.

An approach has been proposed which allows estimating the importance of this relationship. This approach contains a method to measure the geographical distribution of linguistic features. Beyond the particular context of

predicting language vitality, this method constitutes a promising tool to unravel mechanisms that control the spatial distribution of language features.

The proposed approach was implemented with most current linguistic and geographical data: *Jaziky mira* [Academia and Indrik, 1993-2010], *World Language Mapping System* [GMI and SIL , 2010] and the *Atlas of the World's Languages in Danger* [Moseley, 2009].

Besides, it has been outlined how these three very recent databases can be joined in order to conduct quantitative linguistic studies when geographic parameters are involved.

Future work could contain the aspects listed in the following. Despite the fact that the presented methods seem to be quite robust to the buffer size, it would be valuable to study it. Further evaluation of the quality and reliability of the indices M and C would be beneficial. Professor Pablo Jensen's results from ongoing research could be of high value for this purpose[1]. As well, it should be researched whether the temporal stability of linguistic features plays a role for the measurement of their interactions. Following that, it may be promising to further explore alternatives to the complete aggregation of the location qualities of the features manifested by a linguistic community when predicting its vitality. Finally, to predict language vitality more precisely, further investigation of the role and relevance of language similarity in the context of language competition seems useful. A definition of language similarity based on typological features *which is adapted to the particular context of language competition* could constitute a helpful tool for this purpose.

Bibliography

[Abrams and Strogatz, 2003] D. M. Abrams and S. H. Strogatz. Linguistics: Modeling the dynamics of language death. Nature 424(6951), 900–900. 2003. [Online]. Available: http://dx.doi.org/10.1038/424900a

[Academia and Indrik, 1993-2010] Jazyki mira (Languages of the World). Moscow: Academia and Indrik, 1993-2010. [Online]. Available: http://ww.dblang.ru/en

[Castelló, Eguíluz, and Miguel, 2006] X. Castelló, V. M. Eguíluz, and M. San Miguel. Ordering dynamics with two non-excluding options: bilingualism in language competition. New J. Phys. 8(12), 308. 2006. [Online]. Available: http://dx.doi.org/10.1088/1367-2630/8/12/308

[1] Professor Jensen (French National Center for Scientific Research CNRS, France) has published some of his most recent results on his homepage: http://perso.ens-lyon.fr/pablo.jensen/.

[Castelló et al., 2007] X. Castelló, L. Loureiro-Porto, V. M. Eguíluz, and M. San Miguel. The Fate of Bilingualism in a Model of Language Competition. In: Advancing Social Simulation: The First World Congress, 83-94. Eds. S. Takahashi, D. Sallach, and J. Rouchier. Japan: Springer, 2007. [Online]. Available: http://dx.doi.org/10.1007/978-4-431-73167-2_9

[GMI and SIL, 2010] World Language Mapping System. Global Mapping International and SIL International, 2010. [Online]. Available: http://www.gmi.org/wlms

[Jensen, 2006] P. Jensen. Network-based predictions of retail store commercial categories and optimal locations. Phys. Rev. E 74(3), 035101(R). 2006. [Online]. Available: http://dx.doi.org/10.1103/PhysRevE.74.035101

[Jensen, 2009] P. Jensen. Analyzing the Localization of Retail Stores with Complex Systems Tools. In: Advances in Intelligent Data Analysis VIII: 8th International Symposium on Intelligent Data Analysis, Lecture Notes in Computer Science, 5772/2009, 10–20. Eds. N. M. Adams, C. Robardet, A. Siebes, and J-F. Boulicaut. Berlin Heidelberg: Springer-Verlag, 2009. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-03915-7_2

[Lewis, 2009] P. M. Lewis (Ed.). Ethnologue: Languages of the World, Sixteenth edition. Dallas, Texas: SIL International, 2009. [Online]. Available: http://www.ethnologue.com

[Mackey, 2001] W.F. Mackey. The ecology of language shift. In: The Ecolinguistics Reader: Language, Ecology and Environment, 67-74. Eds. A. Fill and P. Mühlhäusler. London: Continuum International Publishing Group, 2001.

[Mira and Paredes, 2005] J. Mira and A. Paredes. Interlinguistic similarity and language death dynamics. Europhys. Lett. 69(6), 1031–1034, 2005. [Online]. Available: http://dx.doi.org/10.1209/epl/i2004-10438-4

[Monteagudo and Santamarina, 1993] H. Monteagudo and A. Santamarina. Galician and Castilian in contact: historical, social, and linguistic aspects. In: Trends in romance linguistics and philology 5, 117–174. Eds. J. Green and R. Posner. Paris: Walter de Gruyter, 1993.

[Moseley, 2009] C. Moseley (Ed.). Atlas of the World's Languages in Danger. Unesco, 2009. [Online]. Available: http://www.unesco.org/culture/en/endangeredlanguages/atlas

[Palmer, 1997] S. Palmer. Language of work: the critical link between economic change and language shift. In: Teaching indigenous languages, 263-286. Ed. J. Reyhner. Flagstaff, Arizona: Northern Arizona University, 1997.

[Patriarca and Heinsalu, 2009] M. Patriarca and E. Heinsalu. Influence of geography on language competition. Physica A: Statistical Mechanics and its Applications 388(2-3), 174–186. 2009. [Online]. Available: http://dx.doi.org/10.1016/j.physa.2008.09.034

[Patriarca and Leppänen, 2004] M. Patriarca and T. Leppänen. Modeling language competition. Physica A: Statistical Mechanics and its Applications 338(1-2), 296–299. 2004. [Online]. Available: http://dx.doi.org/10.1016/j.physa.2004.02.056

[Posner and Rogers, 1993] R. Posner and K. H. Rogers. Bilingualism and language conflict in Rhaeto-Romance. In: Trends in romance linguistics and philology 5, 117–174. Eds. J. Green and R. Posner. Paris: Walter de Gruyter, 1993.

[Schulze and Stauffer, 2006] C. Schulze and D. Stauffer. Monte Carlo simulation of survival for minority languages. In: Advances in Complex Systems (ACS) 9(3), 183–191. Ed. F. Schweitzer. Zurich: Frank Schweitzer, 2006. [Online]. Available: http://dx.doi.org/10.1142/S0219525906000719

[Sutherland, 2003] W. J. Sutherland. Parallel extinction risk and global distribution of languages and species. Nature 423(6937), 276–279. 2003. [Online]. Available: http://dx.doi.org/10.1038/nature01607

[Teşileanu and Meyer-Ortmanns, 2006] T. Teşileanu and H. Meyer-Ortmanns. Competition of Languages and Their Hamming Distance. International Journal of Modern Physics C 17(2), 259–278. 2006. [Online]. Available: http://dx.doi.org/10.1142/S0129183106008765

[Unesco, 2010] www.unesco.org/culture/en/endangeredlanguages. Safeguarding endangered languages. Unesco, 15.03.2010.

[Wang and Minett, 2005] W. S-Y. Wang and J. M. Minett. The invasion of language: emergence, change and death. Trends in Ecology & Evolution 20(5), 263–269. 2005. [Online]. Available: http://dx.doi.org/10.1016/j.tree.2005.03.001

[Wardhaugh, 1987] R. Wardhaugh. Languages in competition: dominance, diversity, and decline. Oxford: Blackwell, 1987.

Author's Information

Samuel F. Omlin *– Section of Information Technologies and Mathematical Methods, University of Lausanne, CH-1015, Lausanne, Switzerland; e-mail: [firstname].[familyname]@unil.ch*

Major Fields of Scientific Research: Applications of Spatial Statistics, Spatial Analysis and Geographical Information Systems in Linguistics, Language Dynamics, Textual Statistics, Natural Language Processing, Supercomputing.