Int. J. Appl. Math. Comput. Sci., , Vol. , No. , –
DOI:

EVOLUTIONARY LANGUAGE COMPETITION - AN AGENT-BASED MODEL

Michał Grzejdziak [a]

[a]Faculty of Mathematics and Computer Science, Jagiellonian University, Łojasiewicza 6, 30-348 Kraków, Poland

In this paper we consider two-language competition in the context of evolutionary linguistics and propose an agent-based model of this phenomenon based on the concept of naming game. We treat language as a complex adaptive system, which evolves under the influence of local interactions in a population of agents. In many cases we obtain a final scenario in which one language totally extincts. However, under certain conditions, the proposed model predicts the emergence of a new language which is a ”mixture” of both competing languages. We compare obtained results with the results of another model of the phenomenon.

Keywords: language competition, agent-based models, evolutionary linguistics

1. Introduction

1.1. Language competition. Over 40% of around 6000 world languages are threatened with extinction within next 100 years, according to alarming UNESCO data from 2010 (Moseley, 2010). Contemporary escalation of the problem is explained by a phenomenom of language competition (Abrams and Strogatz, 2003; Patriarca and Leppännen, 2004; Castelló et al., 2008). It is assumed that languages in a given area ”compete” with each other for users, what results in the ”winning” of the ”strongest” one. The discussion on the topic was started in the paper by D. M. Abrams and S. H. Strogatz (2003), in which the authors described a model of language competition in the form of a differential equation. The model assumed that members of a multilingual population decide which language to use, on the basis of its relative status, defined as ”a parameter that reflects the social or economic opportunities afforded to its speakers” (Abrams and Strogatz, 2003). The authors obtained very similar results to real data concerning a few contemporary endangered languages. Further papers on the phenomenon proposed more complex models, which considered geographical distribution of a population (Patriarca and Leppännen, 2004) or the role of social structure and bilingualism (Castelló et al., 2008). For simplicity, all mentioned research treated language as fixed, while it is generally regarded as a complex adaptive system (Lipowska, 2016). In this paper, in contrast, we propose a model of language competition which takes into account its complexity and adaptivity.

1.2. Evolutionary approach. The problem of language competition can be described by means of evolutionary linguistics methods. It can be formalized on the grounds of applied evolutionary epistemology, a theory developed by Nathalie Gontier (2017), who emphasises the necessity of identification of language evolution units, levels and mechanisms, which are defined as below (Gontier, 2017):

1. ”X is a unit if one can minimally point out one level where X evolves, and one mechanism whereby X evolves” (*What evolves?*).

2. ”X is a level if one can minimally point out one unit that evolves by minimally one mechanism at X” (*Where does it evolve?*).

3. ”X is a mechanism if one can minimally point out one unit that evolves at one level by means of X” (*How does it evolve?*).

We will define the language competition phenomenon as follows: in an area there is a population, in which each member at the initial moment uses one of two different languages. Then, as a result of local interactions between members of population, both languages evolve. Now, we can take competing languages as units, population as level and local interactions as mechanisms of language evolution. Furthermore, we can identify the language competition phenomenon as language coevolution. Comprehensive approach to the problem would demand an analysis of dynamics and correlations of all elements of language (morphology, syntactics, phonology, etc.), which is probably impossible to include in a single model. For this reason, in this paper, we will limit the problem to the coevolution of lexical systems (lexicons), assuming that other elements of language have only negligible impact on it. Therefore, in the proposed model we will examine lexicons as units of language evolution.

1.3. Motivation. Conducting research on this problem can help providing answers to problems in multiple disciplines. Primarily, it can suggest some solutions in the design of language-preserving programmes (Abrams and Strogatz, 2003). Also, it can be helpful in social sciences and humanities. For example, the study of history knows many cases of language competition. One of them was the extinction of East Germanic languages (Visigothic, Ostrogothic or Vandalic) (Moulton and Buccini, 2014) in the Middle Ages. These languages, despite their dominating status as the language of rulers, ”lost” a language competition against Latin and Roman languages.

The proposed model can be helpful in understanding such cases. Mainly, the following questions will be analysed:

- How does language competition proceed, depending on initial conditions?
- Can a new language, being a mixture of competing languages, emerge? If yes, under what conditions?
- What influence on a language competition does an imposed language have? We will understand an imposed language as a language dominating in some areas of life in a population: e.g. Latin in medieval church, official language in a state or English in contemporary culture and science.
- What influence on language competition do some mechanisms of language preservation have?

2. Model description

2.1. Naming games. The proposed model is based on an idea of a tool frequently used in evolutionary linguistics: naming game in the agent-based model (Steels, 2011). It can be shortly described as follows: each unit of a given population is identified with a part of a computer program, which is called an agent. Each agent has a set of object - list of words pairs, which is called a lexicon. Agents interact pairwise in naming games, during which one agent is called a speaker, the other one a listener. The speaker chooses one object and communicates one of the words from the corresponding list to the listener. If the listener has this word in the corresponding list in its lexicon, we call the interaction is ”successful”, otherwise we call the interaction ”unsuccessful”. Depending on the result, the agents modify their lexicons following specifically defined rules.

In the majority of such models the interactions concerning different objects are independent from each other. Therefore, apart from one simulation in an n-object environment, n independent simulations in an one-object environment can be performed. However, in the context of language competition, such an assumption seems to be too simplifying. Competing languages can vary a lot in their perception of the world, i.e. the sets of objects recognized by them can be very different.

2.2. Agents and languages. In the proposed model, there is a population (a set) of $N = n^2$ agents in a d-object environment. Agents have lexicons, which are sets of object - list of words pairs. Words and objects are represented by integer values. Each word has assigned weight, a number from $[0.0; 3.0]$ range, which represents a degree of acquaintance with it. Weight 0.0 states here for complete unawareness of a word, weights other than 0.0 for higher degrees of acquaintance. Weight 3.0 denotes complete acquaintance with a word. Each word belongs to one of two languages (sets of words), which we denote by ℓ_1 (language 1) and ℓ_2 (language 2). A word belongs to ℓ_1 if its integer value is odd; otherwise it belongs to ℓ_2. In a lexicon of an agent X, a word of the highest weight of those in a list corresponding to an object will be called a dominating word of this object. On this basis we can precise the notion of dominating language of the agent X. Let $dom_i(X)$ denote the total number of dominating words from ℓ_i in the X's lexicon and $sum_i(X)$ denote the sum of weights of all words from ℓ_i in the X's lexicon (for $i = 1, 2$). The following criteria will determine the dominating language of X:

- If $dom_1(X) > dom_2(X)$, then ℓ_1 is the dominating language of X.
- If $dom_1(X) < dom_2(X)$, then ℓ_2 is the dominating language of X.
- If $dom_1(X) = dom_2(X)$ and $sum_1(X) > sum_2(X)$, then ℓ_1 is the dominating language of X.
- If $dom_1(X) = dom_2(X)$ i $sum_1(X) < sum_2(X)$, then ℓ_2 is the dominating language of X.
- If none of the above occurs, then ℓ_1 is the dominating language of X (it is chosen arbitrarily, but such a situation is very unlikely in the model).

An agent of dominating language ℓ_i. will be called i-lingual. Let P denote a given population. We define the following characteristics of languages:

- A population of ℓ_i will be a set $P_i \subset P$ of all i-lingual agents. Its size will be denoted by N_i.

- An environment of ℓ_i will be a set of all objects recognized with a weight greater than 0.0 with a word belonging to ℓ_i in any agent's lexicon. Its size will be denoted by d_i. The ratio $\frac{d_i}{d}$ will be called a degree of development of ℓ_i.

- A spread of ℓ_i will be the ratio

$$\frac{\sum\limits_{X\in P} sum_i(X)}{3.0Nd}. \qquad (1)$$

In other words, it will be the ratio of the sum of weights of all words belonging to ℓ_i in any agent's lexicon to the maximal sum of weights in a monolingual population. It will be denoted by s_i.

Obviously, in any moment we have:

$$N_1 + N_2 = N \qquad (2)$$

and

$$\forall_{i=1,2}\, d_i \leqslant d \qquad (3)$$

To represent geographical distribution of the population, agents are placed in the vertices of a weighted $n \times n$ lattice graph of edges of weights $1.0 + \epsilon$, complemented with edges of weights ϵ to a complete graph. Vertices are counted from the first row and the first column: vertex i is the vertex of coordinates $(\lfloor \frac{i}{n} \rfloor, i \pmod n)$. The weight w_{ij} of the edge connecting vertex i with vertex j is given by:

$$w_{ij} = \begin{cases} 1.0 + \epsilon & \text{if } i = j \pm 1 \text{ or } i = j \pm n \\ \epsilon & \text{otherwise} \end{cases} \qquad (4)$$

Initialization of the agents and their lexicons is determined by the parameters N_1^0, N_2^0, d_1^0 and d_2^0. For both languages, N_i^0 agents are initialized as i-lingual, with the lexicons containing one word for d_i^0 objects; an object j $(j = 1, 2, \ldots, d_i^0)$ is initialized with a word of value $2 * (j - 1) + i$ and weight 3.0. A sample initialization of lexicons is shown in the tables 1 and 2. 1-lingual agents are initially placed in the vertices of the indexes from 1 to N_1^0, 2-lingual agents are placed in the remaining vertices. A sample initialization of the structure is shown in the figure 1.

Object	Word	Weight
1	1	3.0
2	3	3.0
3	5	3.0
4		

Table 1. 1-lingual agent initialization for $d = 4$, $d_1^0 = 3$

Object	Word	Weight
1	2	3.0
2	4	3.0
3		
4		

Table 2. 2-lingual agent initialization for $d = 4$, $d_2^0 = 2$

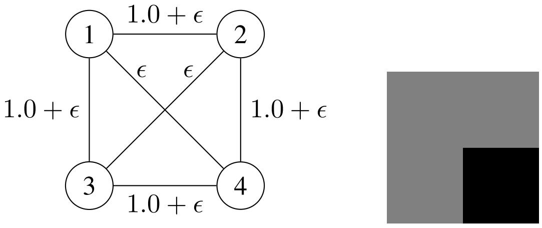


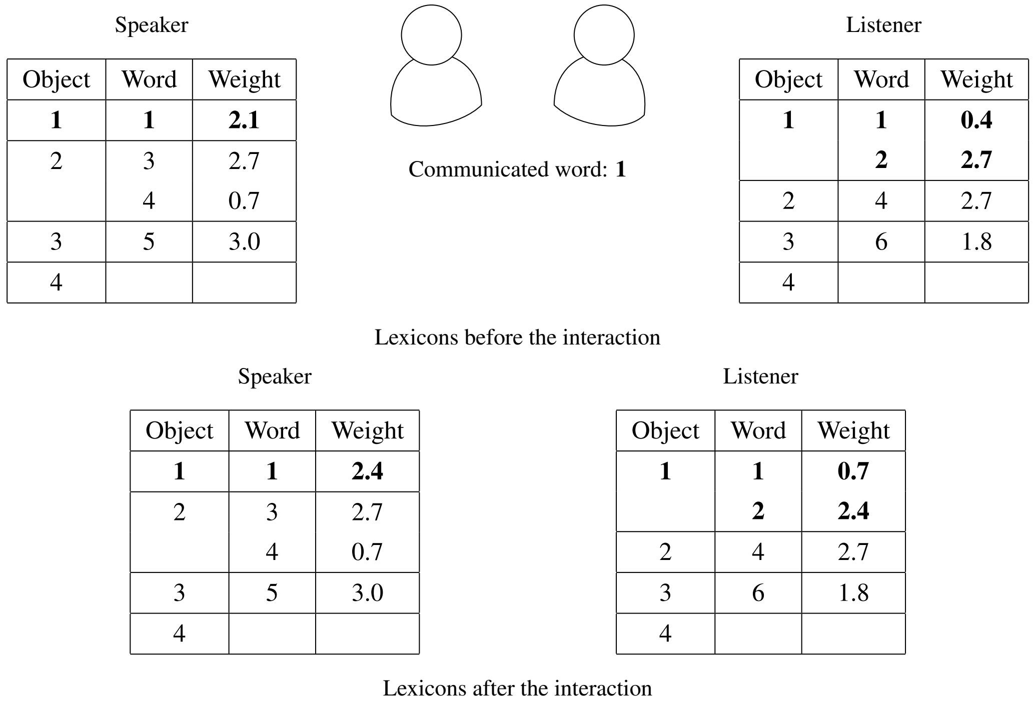
Fig. 1. Initialization of the population structure given by lattice graph for $N = 4$, $N_1^0 = 3$, $N_2^0 = 1$. 1-lingual agents are coloured grey, 2-lingual agents are coloured black

2.3. Interactions between agents. Interactions between agents are as follows: at first a number $x = 1, \ldots, N$ is randomly generated[1] and the agent x (i.e. agent in the vertex x) becomes a speaker. Then, a listener is randomly selected. The probability p_{xy} of selecting agent y as a listener is given by:

$$p_{xy} = \frac{w_{xy}}{\sum\limits_{k=1}^{N} w_{xk}} \qquad (5)$$

The speaker randomly chooses number $o = 1, \ldots, d$. If it has in its lexicon at least one word corresponding to the object o, then it communicates to the listener one such a word of the greatest weight. Otherwise, it creates a new corresponding word in the following way: if it is i-lingual, it randomly generates an integer number k and the new word receives value $2k + i$ and weight 1.0, and is communicated to the listener. If the listener recognizes the communicated word, then the game is successful, otherwise it is unsuccessful. Agents keep a history of successful interactions for both languages and on this basis modify their lexicons after each interaction. If the interaction is successful, then both agents increase the weights of the communicated word by 0.3 multiplied by the ratio of the number of successful interactions in

[1] We generate random numbers with uniform distribution



Speaker

Object	Word	Weight
1	**1**	**2.1**
2	3	2.7
	4	0.7
3	5	3.0
4		

Communicated word: **1**

Listener

Object	Word	Weight
1	**1**	**0.4**
	2	**2.7**
2	4	2.7
3	6	1.8
4		

Lexicons before the interaction

Speaker

Object	Word	Weight
1	**1**	**2.4**
2	3	2.7
	4	0.7
3	5	3.0
4		

Listener

Object	Word	Weight
1	**1**	**0.7**
	2	**2.4**
2	4	2.7
3	6	1.8
4		

Lexicons after the interaction

Fig. 2. A sample successful interaction, for $d = 4$

the corresponding language, in which they participated, to the total number of interactions in this language in which they participated. They also decrease by the same value weights of all other words corresponding to the object o in their lexicons. Such defined behaviour after successful interactions is reflecting the assumption, that during language competition agents prefer using more successful language. Value 0.3 is adopted conventionally; by that means it is necessary for an agent to interact in at least 10 successful interactions concerning one word to increase its weight by maximal available value 3.0. If the interaction is unsuccessful, the listener acquires the communicated word with the weight 1.0 and the speker decreases its weight by 0.3. When the weight of a word decreases to 0.0 in an agent's lexicon, then the agent removes it from the lexicon, unless it is the only word in its lexicon corresponding to an object; if so, it is left with conventional weight 0.1. It is assumed here, that the process of acquiring word is more effective than the process of learning, and that the only word denoting an object cannot be forgotten. A sample interaction between agents is shown in the figures 2. For simplicity, for both agents respective ratios of successess to all interactions are assumed 1.0.

2.4. Model variants. The above rules describe the basic variant of the proposed model. In addition, two other variants will be analysed: one with a "total speaker", the other one with a "total listener". Both include, before each simulation step[2], v additional interactions of agents from a population with respectively total speaker or total listener. Total speaker has its own lexicon, which contains one word belonging to the variant language ℓ_{i_v} for $d^0_{i_v}$ objects; it plays a role of a speaker in each interaction it participates in. It can reflect an influnece of an imposed language in a population. Total listener plays a role of a listener in each interaction it participates in; each such interaction is successful. It can reflect some machanisms of language preservations. We consider both these variants, together with the machanism of successfull language preference, a substitution for the notion of language status, included in previous research on language competition (Abrams and Strogatz, 2003; Patriarca and Leppännen, 2004; Castelló et al., 2008).

The proposed model can be generalized to any number of competing languages.

[2] a k-th simulation step (or a k-th iteration) states for N subsequently simulated interactions, from $((k-1)N+1)$-th one to kN-th one ($k \in \mathbb{N}_1$)

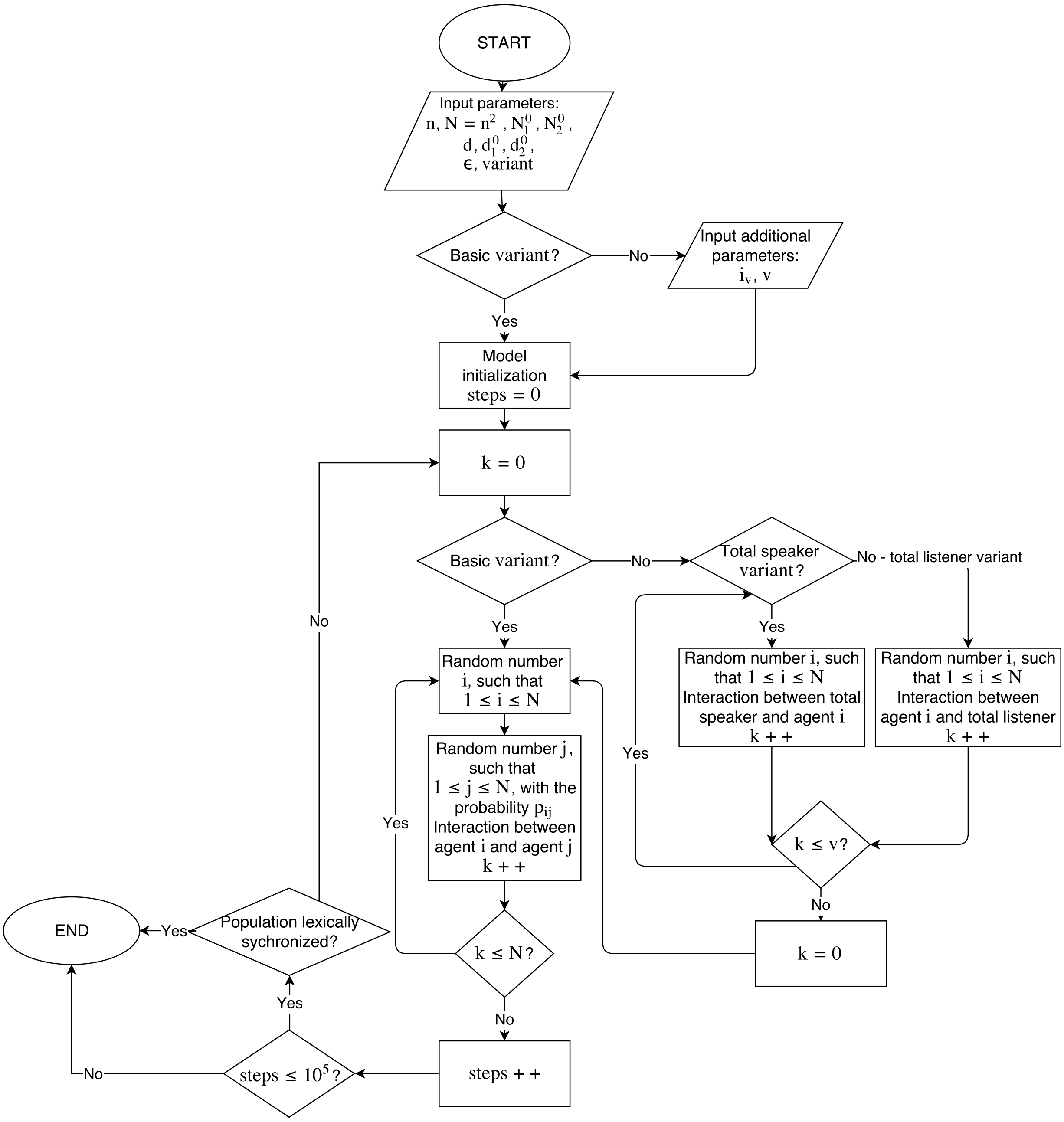

START
Input parameters:
n, N = n², N_1^0, N_2^0,
d, d_1^0, d_2^0,
ϵ, variant
Basic variant?
No
Input additional parameters:
i_v, v
Yes
Model initialization
steps = 0
k = 0
Basic variant?
No
Total speaker variant?
No - total listener variant
Yes
Random number i, such that 1 ≤ i ≤ N
Random number i, such that 1 ≤ i ≤ N
Interaction between total speaker and agent i
k + +
Random number i, such that 1 ≤ i ≤ N
Interaction between agent i and total listener
k + +
Random number j, such that
1 ≤ j ≤ N, with the probability p_{ij}
Interaction between agent i and agent j
k + +
k ≤ v?
No
k = 0
No
Yes
Yes
END
Yes
Population lexically sychronized?
k ≤ N?
No
steps + +
steps ≤ 10⁵?
No

Fig. 3. A flowchart of model simulations

3. Model simulations

The set of paramaters of the model include:

- parameter n determining the population size N, together with initial sizes N_1^0 and N_2^0 of both languages populations

- a number d of all objects and initial environments' sizes d_1^0 and d_2^0 of both languages

- ϵ

- model variant: basic, with total speaker or with total listener (together with the variant language parameter i_v and the parameter v)

For the model examination, computer simulations were performed in a program implemented in Java. Each of them was performed in 10^5 steps but was terminated when a population of agents achieved the state of lexical synchronization (lexicons identity with an accuracy of weights of words). The figure 3 shows a flowchart of model simulations.

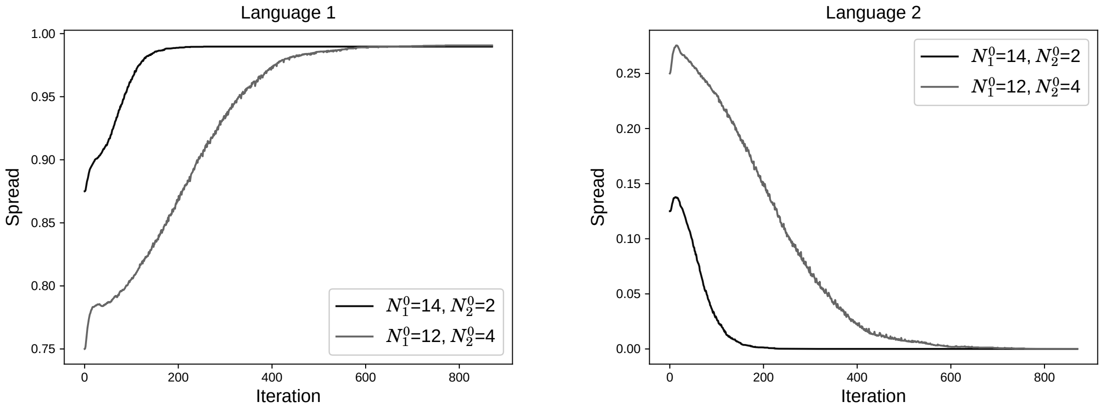


Fig. 4. The dynamics of spreads of competing languages for different sizes of initial language populations

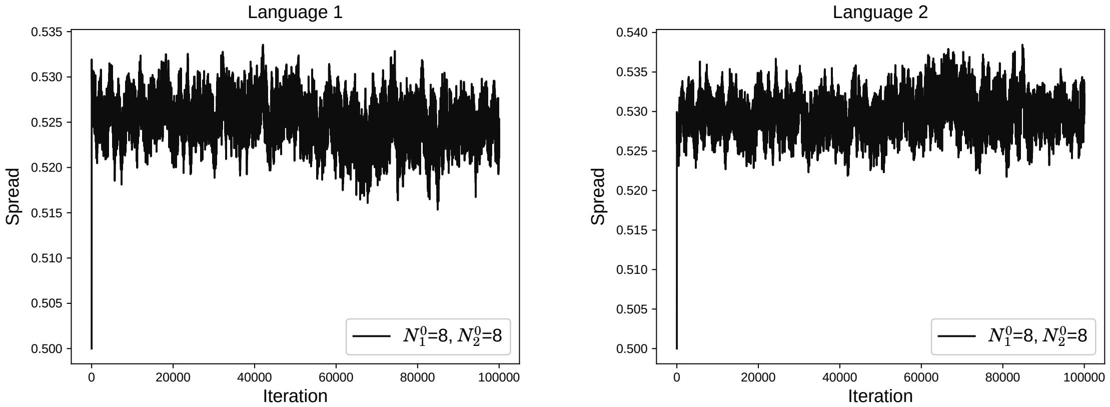


Fig. 5. The dynamics of spreads of competing languages for equal sizes of initial language populations

All simulations analyzed here were performed with $N = 16$, $d = 5$, $\epsilon = 0.01$. For each examined set of parameters, at least 100 independent simulations were taken, and their results were averaged. During simulations the dynamics of characteristics of both languages: their degrees of development, spreads and numbers of users were measured.

4. Experiments

4.1. Initial populations' sizes' impact. In order to examine the model behaviour depending on the parameters N_1^0 and N_2^0, simulations for four different sets of parameters were perforemed. The dynamics of spreads of competing languages obtained is shown on the plots in the figure 4. Conclusions can be drawn, that any disproportion in initial sizes of language populations results in total extinction of the initially outnumbered language, and agents' compromise lexicon consists only of words from the language of greater initial population. It is worth noting, that the less disproportion in initial sizes is, the longer language competition lasts: for the ratio 12.5% : 87.5% of sizes competition lasts 141 iterations on average, for 25% : 75% it lasts 454 iterations, and for 37.5% : 62.5% - 6 758 iterations. Therefore, although the final state is identical in all cases, the time to reach it is very sensitive to changes in initial ratio of language populations sizes. The results obtained are similar to the actual situations, e. g. to the mentioned earlier extinction of East Germanic languages (Moulton and Buccini, 2014). For the ratio 50% : 50% population didn't reach a language compromise in any simulation. The plot of the spreads of competing languages is shown in the figure 5. The spreads of both languages oscillate around 53%, what probably means languages stabilisation in initial populations with constant penetration by both languages of a border between populations. The penetration is performed only to a small extent, and none of languages can get an advantage. Population of agents reaches an equilibrium, in which three regions can be distinguished: one of ℓ_1, one of ℓ_2. and one bilingual on the border.

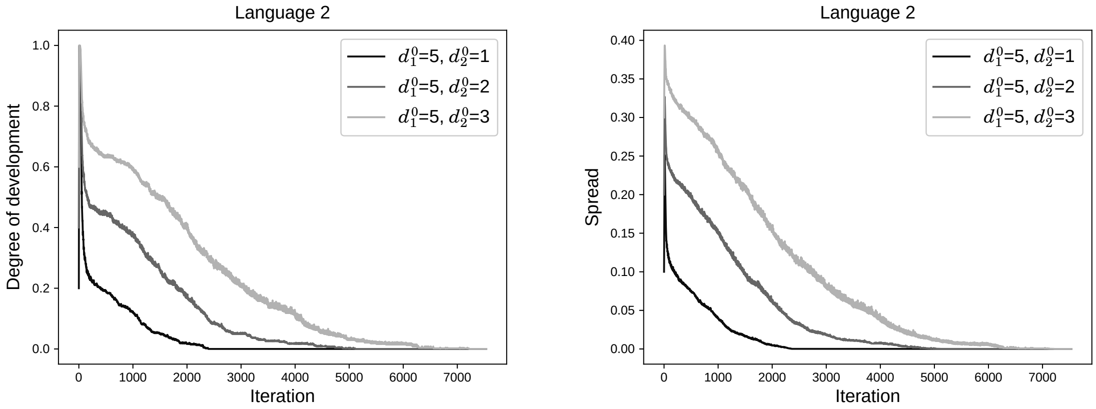


Fig. 6. The dynamics of the degree of development and the spread of the language initially less developed for simulations with equipotent initial language populations.

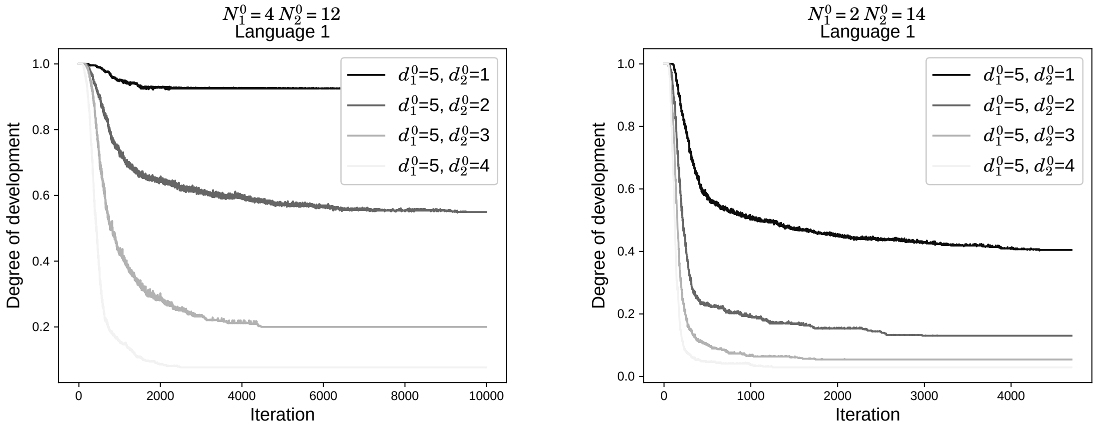


Fig. 7. The dynamics of degrees of development of ℓ_1 for different initial population sizes and different initial environments' sizes.

4.2. Initial environments' sizes impact. The influence of initial environments' sizes d_1^0 and d_2^0 was examined in simulations with equipotent initial populations. The results are shown in the figure 6. The plots show that, in general, the language of the smaller initial environment's size totally extincts. Therefore, a disproportion in initial environments' sizes has similar effect to a dispropotion in initial sizes of language populations. An essential difference can be observed between times of reaching lexical synchronization. In case of decreasing disproportion in initial environments' sizes - similarly to disproportion in sizes of initial language populations - an increase in time of reaching language comrpomise can be observed, but it is much less sensitive to changes in the parameter: for d_1^0, d_2^0 being respectively 5 and 1, the competition lasted 1 267 iterations on average, for 5 and 2 - 2 562 iterations, 5 and 3 - 3 454 iterations. A slightly different behaviour can be observed for $d_1^0 = 5$ and $d_2^0 = 4$; in 80% performed simulations the language initially less developed totally extincts after 6 824 on average, in the other 20% simulations populations reaches an equilibrium similar to this observed in case $N_1^0 = N_2^0$. Probably, it is a result of indeterministic time of finding a compromise word for an initially unknown by the language less developed object; in some cases it happens so quickly, that the other language doesn't manage to get an advantage. However, in most cases it happens too slowly.

4.3. Joint impact of initial populations' sizes and initial environments' sizes. Simulations performed for different sets of parameters has shown, as expected, that if one language has initially both greater population's size and environment's size, then the other one totally extincts in competition; the compromise lexicon consists only of the words from the language initially in advantage. Interesting results can be observed in a situation, when initially one language has an advantage in population's size, and the other in environment's size. Some examined cases are shown in the figure 7. It appears,

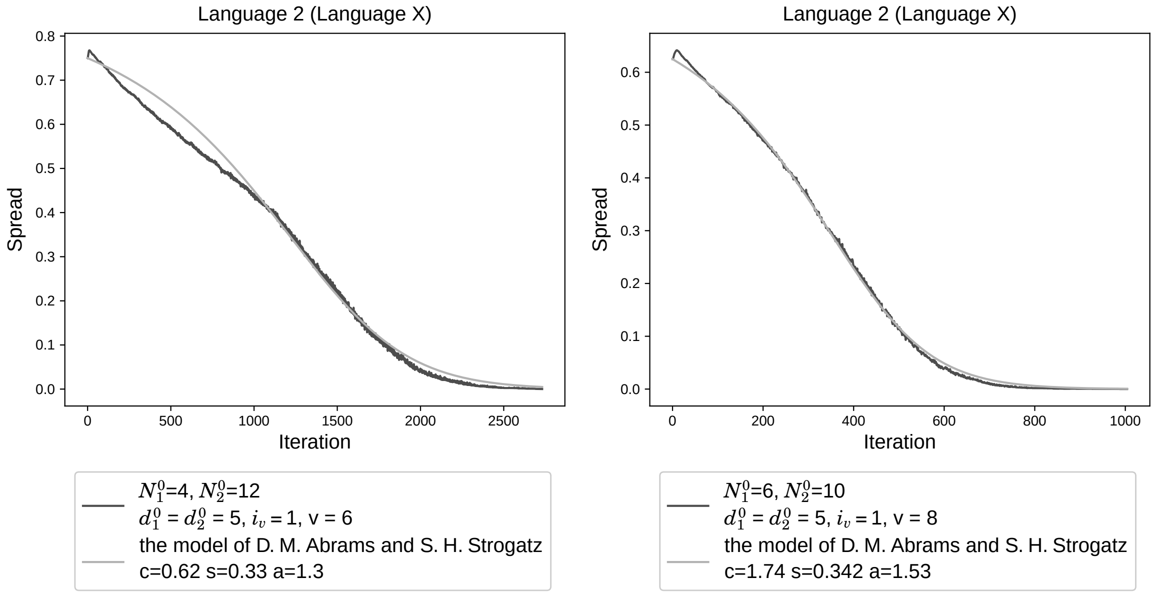


Fig. 8. A comparison of the proposed model and the model of D. M. Abrams and S. H. Strogatz (simulated numerically with $x(k+1) = x(k) + dt\frac{dx(k)}{dt}$ for $k \in \mathbb{N}$, $dt = 0.01$). In both cases, ”total speaker” uses ℓ_1

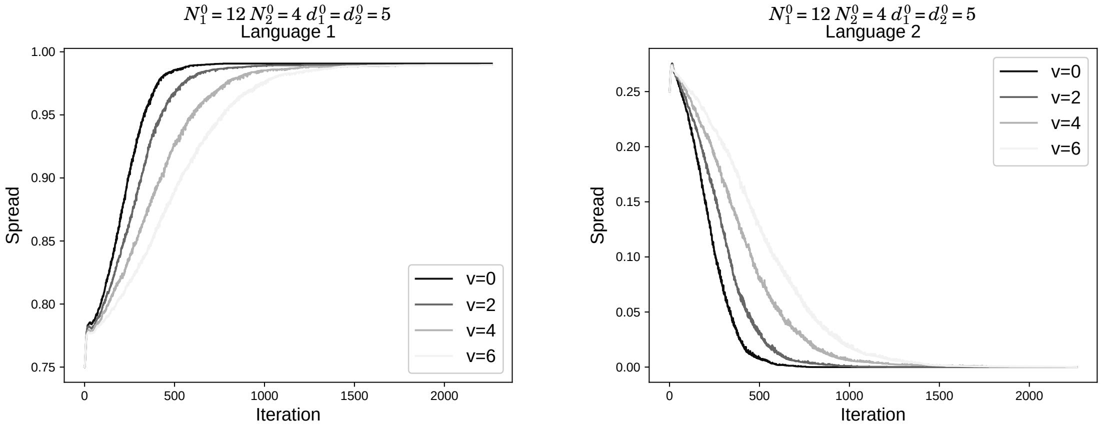


Fig. 9. Spread of langauges for the simulations in the ”total listener” variant

that in the proposed model, in such situations, the compromise lexicon can be based on both languages. Such behaviour could not appear in previous models of language competition (Abrams and Strogatz, 2003; Patriarca and Leppännen, 2004; Castelló et al., 2008). In the proposed model, it occurs for relatively big disproportion in populations sizes (at least 25% : 75%) and relatively big disproportion in environments' sizes (for the ratio of population sizes 25% : 75% it is 5:2, for 12.5% : 87.5% - 5:1). Simulations with a smaller disproportion in populations sizes resulted in a total extinction of initially less developed language. This phenomenon of language mixture, appearing in the model, can be observed also in reality: an example here could be language competition between Latin and European ethnic languages, which resulted in a lot of loanwords in them. In this context, derivation of Creole languages can also be considered.

4.4. ”Total speaker” impact. The ”total speaker” variant in the model allows to examine the notion of language ”status”. It was considered by D. M. Abrams and S. H. Strogatz (2003) as a parameter of their model in the form of the following differential equation:

$$\frac{dx}{dt} = yP_{yx}(x,s) + xP_{xy}(x,s) \tag{6}$$

where x and $y = 1 - x$ denote percentile numbers of users of languages X and Y, and P_{yx} denotes an average probability of language change from to language X by users of langauge Y. The above equation can be expressed as

$$\frac{dx}{dt} = ycx^a s + xcy^a(1-s) \tag{7}$$

where s states for a relative "status" of language X to language Y, and c and a are constants. For known cases, the value of a was found roughly constant and equal to 1.31 ± 0.25 (where 0.25 states for a standard deviation). The status parameter was described as "a parameter that reflects the social or economic opportunities afforded to its speakers" (Abrams and Strogatz, 2003), and its value is adjusted to statistical data. However, this approach offers only little explanatory value; appropriate values of the status parameter can be adjusted, but it is hard to determine, what they could mean. Simulations performed in the proposed model has shown, that the "total speaker" variant can give similar results to the model proposed by D. M. Abrams and S. H. Strogatz (2003). A comparison of both models is shown in the figure 8. The similarity in behaviour of both models indicates, that "total speaker" can be a good representation of language status.

4.5. "Total listener" impact. The "total listener" variant, which can imitate different language-preserving activities, was examined in experiments for different values of variant influence v. The simulations showed, that such activities can extend the time of competition and delay an extinction of an initially outnumbered language. The results of the experiments are shown in the figure 9. The time extension, however, is relatively small: for $v = 0$ an average time of reaching language compromise is 434 iterations, for $v = 2$ it is 539 iterations, $v = 4$ - 728 iterations, and for $v = 6$ - 925 iterations. In the last case, language competition lasts on average slightly over two times longer than in the case of $v = 0$.

5. Conclusions

After a partial verification of the proposed model with actual examples and comparison to the acclaimed model of D. M. Abrams and S. H. Strogatz (2003) we can conclude, that the proposed model can represent the phenomenon of language competition. Therefore, it can support research in different disciplines, helping in veryfing hypothesis concerning the problem. It can also support planning of language-preserving activities. We believe, that evolutionary approach applied here can be extended and give more interesting results. For example, other elements of language than lexical can be taken into consideration. Potentially, the proposed model could be extended in order to examine better the issue of Creole languages derivation.

Acknowledgment

The author wishes to thank Marek Skomorowski for his support and valuable suggestions.

References

Abrams, D. M. and Strogatz, S. H. (2003). Modelling the dynamics of language death, *Nature* **424**: 900. 10.1038/424900a

Castelló, X. et al. (2008). *Modelling language competition: bilingualism and complex social networks*, World Scientific, pp. 59–66. URL:https://www.worldscientific.com/doi/abs/10.1142/9789812776129_0008 10.1142/9789812776129$_0$008

Gontier, N. (2017). What are the units of language evolution?, *Topoi* **8(4)**: 1–19. 10.1007/s11245-017-9474-8

Lipowska, D. (2016). *Komputerowe modelowanie ewolucji języka (in Polish)*, 1 edn, Wydawnictwo Naukowe UAM, Poznań.

Moseley, C. (Ed.) (2010). *Atlas of the World's Languages in Danger*, 1 edn, UNESCO Publishing, Paris. Accessed on 20 Febr. 2018. URL:http://www.unesco.org/culture/en/endangeredlanguages/atlas

Moulton, W. G. and Buccini, A. F. (2014). East germanic languages. Accessed on 20 Febr. 2018. URL:www.britannica.com/topic/East-Germanic-languages

Patriarca, M. and Leppännen, T. (2004). Modeling language competition, *Physica A* **338**: 296–299. 10.1016/j.physa.2004.02.056

Steels, L. (2011). Modeling the cultural evolution of language, *Physics of Life Reviews* **8**(4): 339–356. URL:http://www.sciencedirect.com/science/article/pii/S1571064511001060 10.1016/j.plrev.2011.10.014

Michał Grzejdziak (born 1997, Warsaw, Poland) is a student of computational mathematics and computer science in the Faculty of Mathematics and Computer Scince of the Jagiellonian University. He is interested in applying computational methods in social sciences and humanities.