DE GRUYTER MOUTON
DOI 10.1515/ijsl-2013-0020 — IJSL 2013; 221: 1–10

Johannes Kabatek and Lucía Loureiro-Porto

Mathematical models meet linguistic data and vice versa: an introduction to this special issue

Johannes Kabatek: University of Tübingen, Germany. E-mail: kabatek@uni-tuebingen.de
Lucía Loureiro-Porto: University of the Balearic Islands, Spain. E-mail: lucia.loureiro@uib.es

Linguistics in the 21st century is faced with a series of methodological innovations that have opened up new ways of describing language: experimental methods allow us to measure the activity of the human brain and to relate this activity to linguistic behavior and ability (Bornkessel-Schlesewsky and Schlesewsky 2009); new discoveries in genetics and in evolutionary biology show how certain genes play crucial roles in our language faculty and allow for a more precise definition of the moment when human language emerged (e.g. Benítez-Burraco et al. 2008); and the analysis of large amounts of data allow for the modeling of linguistic phenomena on a scale impossible to achieve in the past (e.g. Köhler 2012). Surprisingly, many of these innovations have emerged outside the discipline itself. Articles on linguistic issues are now being published in journals such as *Nature*, *Science* and *Physica A* by authors who work in the fields of statistical physics, evolutionary biology, cybernetics and mathematics. The reaction of "genuine" linguists, theoretical and empirical, is often rather skeptical and frequently leads to such studies being viewed with suspicion, or rejected altogether. And indeed, a general tendency in current "scientific" approaches to linguistics is to concentrate principally on the *method* and to impress with sophisticated experiments or quantitative analysis. However methodologically convincing, though, these studies sometimes belie shortcomings of both an empirical and a theoretical nature.

This general observation also holds true for sociolinguistics and for the sociology of language, where a considerable number of recent studies have been published with exciting new proposals on the possibility of modeling individual behavior in a social context and on visualizations of phenomena such as language shift, language change and language death. Of course, as with any other scientific discipline, sociolinguistics has always dealt with models: models of societal stratification, models that link social constellations to linguistic attitudes and linguistic behavior, and even quantitative models that establish correlations

between data collected from a representative number of individuals and which are then projected onto a general sociolinguistic configuration of a society. However, mathematical models have recently been used not only as descriptive instruments but also as productive tools that, based on a few microlinguistic parameters that may be crucial conditioners of human behavior, simulate macrolinguistic processes. Even if some of these studies lack a firm linguistic foundation in one way or another, they nevertheless allow for new insights, and we believe that linguists and sociolinguists should not be blind with respect to the new methods, or ignore what those working in other disciplines are able to contribute to the field. As Blythe and Croft note:

> . . . there are benefits to collaboration between empirical researchers in the humanities, who generally have not been trained in mathematical analysis of quantitative data, and modelers from physics or other strongly mathematical fields, whose models are not always based on empirical data from the domains being modeled. (Blythe and Croft 2010: 19)

Collaborations between technically oriented modelers and linguists should allow for the creation of a heuristic spiral where linguistic data and the experience of linguists serve as a starting point for modeling. The results of modeling should then be critically assessed again, in light of empirical evidence, this in turn leading to the addition of further factors or parameters to the model. Showing this kind of productive interdisciplinary interchange is the main target of the current special issue: to create encounters between mathematical modelers and linguists as a means of improving their particular perspectives on the different phenomena of language contact.

On a more general view, mathematical models have been applied to three major sociolinguistically relevant fields by scholars from other disciplines (mathematics, physics, bioinformatics) working with quantitative data: (a) language competition, (b) linguistic kinship (and typological distances among families), and (c) evolutionary dynamics of languages.

The competition among languages (involving language shift and language death) is one of the most common areas of interest for modelers of dynamic processes, perhaps motivated by the large number of languages which are claimed to be endangered (Crystal 2000). Different phenomena here have been mathematically modeled, including prestige and the importance of the density of speakers (Abrams and Strogatz 2003; Tamariz et al. 2011), the role played by geography (Patriarca and Heinsalu 2009), demographic growth (Zanette 2008; Wichmann and Holman 2008; Kandler 2009), social structure (Gong et al. 2008), complex networks (Liu 2008; Zhou et al. 2008; Castelló et al. 2007, 2008, 2009), the viability of the languages in competition and their possible resilience (Chapel et al. 2010), and the stability of bilingualism (Wickstrom 2005; Mira et al. 2011).

Quantitative studies of kinship between languages and language families have often looked at typological distances. Thus, studies such as de Oliveira et al. (2008) have developed computer-based models that calculate the distribution of language families similar to those assumed in traditional historical and typological studies. Likewise, Petroni and Serva (2010) propose a model for measuring lexical similarities between languages, and Wichmann et al. (2010) evaluate the linguistic resemblance of languages based on normalized Levenshtein distances. In addition, Buch et al. (2011) have recently presented an innovative clustering approach for automated language classification based on an algorithm originally designed for the analysis of similarities between protein sequences.

A third field where mathematical models have been successfully applied is that of the evolutionary dynamics of language. That is to say, issues such as the evolution of grammar, lexis and phonology are accounted for through quantitative models. Thus, for example, grammar has been modeled by Lieberman et al. (2007), whereas both grammar and phonology are assessed in Baxter et al.'s (2009) attempt to evaluate Trudgill's theory on the emergence of New Zealand English as a distinctive dialect (for the measuring of dialect distance and the evolution of language, see also Nerbonne 2010). Finally, lexical issues and the formation of linguistic categories have been the focus of Loreto et al.'s (2008) article on the dynamics described in the Naming Game, a typical experiment carried out in the field of artificial intelligence as a means of observing the emergence of a shared vocabulary among different individuals. Game-theoretic approaches have been used in the modeling of language evolution, with different subtypes of game theory being successfully applied to the description of pragmatically motivated language change phenomena (see Jäger [2008] and the papers in the volume edited by Benz et al. [2011]).

Several of these interdisciplinary studies have been published only in specific sciences journals and have used the kind of scientific language which may not be transparent to linguists. An interesting example in this context is Abrams and Strogatz's (2003) *Nature* paper on language contact and language shift, where, quite artificially, two monolingual communities are brought together, and scenarios of language shift, language death and language survival are "modeled" on the basis of a simple simulation. The data in this research were rather heterogeneous, and their interpretation was somewhat impressionistic, largely in order to make the data fit the model. Both linguists and modelers criticized the approach as too simplistic. A further step was then taken by Minett and Wang (2008), whose model was more sophisticated, and by Castelló et al. (see this issue), who introduce the variable of bilinguals as an important factor in language contact. The models employed in this last collaboration are by no means free of controversy, since they require a degree of simplification which is some-

what difficult to accept by anyone with a background in the humanities (see, for example, the approach to the concept of *volatility* in Castelló et al.'s article, which departs from the idea that all the members in a given society are equally volatile, and which is in fact not commonly witnessed, as noted in Markus and Rozhanskyi's article, this issue). Yet Castelló et al.'s models are based on a fruitful collaboration between linguists and physicists, with a firm basis in both disciplines, and in that sense the study reflects our idea of the heuristic spiral, and allows for further adjustments in future research. In this heuristic spiral, modelers will have to learn more and more about the real complexity of the focus of their activities, in this case language contact. But linguists, sociolinguists and sociologists of language will also have to learn to reduce the enormous complexity of these same objects of study to a few dimensions, thus allowing for successful and useful quantitative modeling. In some ways this need for reduction will be more easily achieved in sociolinguistics than in other linguistic disciplines, since through reduction we effectively move back to the origins of quantitative sociolinguistics in the 1960s and to the early Labovian studies, where large amounts of data on very few features allowed for a coherent projection of a complex sociolinguistic panorama (see the overview in Labov [1994]).

Only a few of these interdisciplinary approaches to the study of linguistic phenomena have been published in linguistics journals (Minett and Wang 2008; de Oliveira 2008; Baxter et al. 2009), hence the fundamental motivation for the publication of this special issue was to bring together scholars from different disciplines and build bridges between experts in modeling and sociolinguistic researchers doing empirical work in different areas. In order to achieve this goal, we organized a workshop on "Modeling language contact: linguistic data and interdisciplinary approaches", which took place at the 43rd Meeting of the *Societas Linguistica Europaea*, held at the University of Vilnius, 2–5 September 2010. The topic of the meeting was "Language contact: at the crossroads of disciplines and frameworks" and, as conveners of the workshop, we thought this would be an ideal setting for scholars from disciplines such as linguistics, computer science and physics to meet and discuss their different perspectives on the study of language contact. The contact established between the different scholars, as well as the resulting discussions, seemed to us a suitable starting point for the preparation of this special issue, which would allow us to bring the discussion on the possibilities and limits of models closer to sociolinguists. As we said, communication among disciplines is absolutely necessary, and we think that the present issue is a good example of how both linguists and modelers are willing to collaborate.

The issue comprises six articles. They are linked by a single thematic thread: an attempt to disentangle the main linguistic and social strategies at work in the

competition between languages in bilingual communities, both from a mathematical and an empirical point of view. The articles discuss mathematical models of the disintegration of languages (de Oliveira et al.) and language competition (Castelló et al.), empirical studies of language competition, whose results constitute feedback for the models themselves (Markus and Rozhanskiy; Jansen), linguistic models of the acquisition of two languages in bilingually raised children (Lleó and Cortés) and empirical studies on the prosodic accommodation of speakers in situations of language contact (Romera and Elordieta).

The opening paper is a mathematical approximation of the disintegration of (proto-)languages, by Paolo Murilo Castro de Oliveira, Adriano O. Sousa and Søren Wichmann. The identification of languages (versus dialects) proves an essential pre-condition for the study of language competition, which has gone largely unnoticed in studies such as Abrams and Strogatz (2003), where it is simply assumed that when two languages compete, one of them dies and the other survives. As we know, present-day languages derive from ancient languages which no longer exist, and in the process of diversification some languages are "stronger" (i.e. more resistant) than others and therefore survive, while others disappear; however, this is not the whole story, since new languages may emerge owing to this language contact situation. De Oliveira et al.'s article constitutes a solid base for the analysis not only of the competition among languages but also the competition among language families, since it pays close attention to the dynamics taking place within the evolutionary trees traditionally used to identify language families. The novelty of the article is the introduction of a technique for the analysis of linguistic phylogenetic trees in which attention is paid to the length of each branch. The model departs from the Swadesh list (1955), which includes words that identify the languages of the world, and branch lengths are calculated on the basis of Levenshtein distances, which basically measure the number of steps necessary for a word X to become a word Y, taking into account that at each step only one of the sounds of the word can be changed. Their main finding is that (proto-)languages disintegrate into new languages at a relatively constant rate, which raises the question of whether the study of language competition should in fact be replaced by the study of competition among entire language groups.

The second article in the issue, by Xavier Castelló, Lucía Loureiro-Porto and Maxi San Miguel, is also inspired by mathematical models of language competition, with special reference to agent-based models. The article constitutes a sound introduction to agent-based models in general and focuses on those that involve the social sciences (cf. Axelrod 2006) before going on to discuss the relevance of complexity theory for the study of language competition. In order to guide the reader, the article offers a careful discussion of central terms from both modeling

and linguistics, such as *prestige*, *network* and *density of speakers*. Since the concept of social network is of greatest importance here, the models presented depart from Abrams and Strogatz's (2003) model, and also from Minett and Wang's (2008) improved version with the introduction of bilingual speakers, and study the different scenarios which obtain for each of the models in different sorts of complex networks: regular lattices, small world networks, fully-connected networks, and networks with community structure. The only parameters taken into account in the models are the prestige of the languages and the volatility of the speakers (i.e. their willingness to shift languages). The article is highly accessible for both modelers and linguists without prior instruction on mathematical approaches to language contact, with results discussed from both a qualitative and a quantitative point of view. Conclusions show that language shift (language competition) can be considered a complex phenomenon in which the interaction among individuals plays a more important role than the actual features of each speaker, which are sometimes over-estimated in the sociology of language, since the only parameters of prestige and volatility account for a wide variety of language contact situations, including language endangerment, language resilience, and language segregation. The conclusions are, nevertheless, tentative, since more detailed quantitative studies of language competition are required before the models can be validated.

The third article in the issue, "Correlation between social and linguistic parameters in modeling language contact: evidence from endangered Finnic varieties" by Elena Markus and Fedor Rozhanskiy, is an empirical study of the competition of Votic and Ingrian, two minority languages in contact in the western part of the Leningrad Oblast of the Russian Federation. These two varieties are only spoken by a small percentage of the population and have coexisted for centuries, which has given rise to different strategies in speakers. While the Ingrians show a higher volatility to adapt their language to Votic, Votes are rather inflexible as regards language change, but less so when it comes to shifting languages. Thus, what the authors observe is that Ingrian and Votic have survived, despite being two minority languages, thanks to the two different kinds of volatility in their respective speakers: while Ingrians are highly volatile at the linguistic level (allowing for language change), Votes are highly volatile at the social level (allowing for language shift). The empirical results of this study constitute interesting information for the future of models such as Castelló et al.'s (this issue), which simply consider that volatility is a property of the system by means of which agents are willing to shift languages.

The fourth article in the issue, "Language maintenance and language loss in marginalized communities: the case of the *bateyes* in the Dominican Republic" by Silke Jansen, constitutes another empirical approach to the dynamics of lan-

guage contact, and evaluates the effects of social aspects such as prestige and the speakers' attitude to both Haitian Creole and Spanish about the survival (or, rather, the death) of Haitian Creole, which is in contact with Spanish. The article explores the data from personal interviews by the author with female immigrants of the second generation of Haitians who moved to the Dominican Republic to work at the *bateyes*, company towns for sugar workers, where Dominicans constitute 50% of the population. Although the informants claim that they are fluent in both Creole and Spanish, the author's subsequent analysis of collected data reveals that Spanish is the dominant language. The higher prestige of Spanish and the stigmatization of Haitian Creole are two of the factors contributing to the attrition and further loss of the latter after the second generation, even though new migrants from Haiti continue to arrive at the *bateyes*. The author, therefore, notes that her data fit with the general assumption in the modeling of language competition that prestige is a powerful factor. Nevertheless, she also calls for the inclusion of further parameters in future models to account for population fluctuation as well as for cross-ethnic networks, since these two features have thus far been absent from the preliminary models designed by mathematical modelers.

The final two articles in the issue pay attention not to language shift, but rather to language change at the prosodic and phonic level, in two language contact situations involving Romance languages. In their article "Modeling the outcome of language contact in the speech of German-Spanish and Catalan-Spanish bilingual children", Conxita Lleó and Susana Cortés carefully analyze the degree of Spanish-German bilingualism of children raised in Hamburg and in Barcelona, as well as Spanish-Catalan bilingualism in Barcelona. The aim of the article is to establish a hierarchy of factors to help predict the phonological competence of bilingually-raised children and, therefore, to serve as the base for a linguistic model of bilingualism. By studying two groups of bilinguals (on the one side, children with a Spanish parent being raised in Hamburg, and children with a German parent being raised in Madrid; on the other, children being raised bilingually in Spanish and Catalan in Barcelona), the authors study the effects of bilingualism on the production of phonological features. Bilingualism may cause the acceleration of the production of a given phonological variable, but it may also delay it, at the same time as it can also be transferred from one language to another. By measuring each of the variables and the factors that might play a role in their acceleration, delay or transfer, the authors evolve two hierarchies of these factors, one accounting for internal factors (such as the frequency of the variable), and one which accounts for external factors (such as the dominant language in the district). The authors' conclusions are that the outcome of bilingualism can be predicted following these hierarchies, since the frequency of a phonological variable plays a more important role than the markedness or the

uniformity of the variable. Likewise, taking into account the social context, the language which is dominant in the district plays a more important role in the child's acquisition than the language spoken with their mates or the language spoken by the parents to each other.

The last article in the issue is Magdalena Romera and Gorka Elordieta's "Prosodic accommodation in language contact: Spanish intonation in Majorca", in which they study the contact between Majorcan Catalan, Majorcan Spanish and Peninsular Spanish with the aim of determining whether Spanish native speakers who arrive in Majorca are more heavily influenced by Majorcan Catalan or by the variety of Spanish spoken in Majorca. By designing an experiment in which the authors hold a semi-directed conversation with native Spanish speakers who arrived in Majorca no more than eight years before the experiment took place, they observe that these speakers consistently accommodate the prosodic features of Majorcan Catalan (especially in interrogative sentences). Even if the possibility that Majorcan Catalan may influence the Spanish variety of these native speakers is not ruled out, the authors conclude that, given that none of these speakers are fluent in Catalan, the accommodation observed in their prosody is caused by the contact with Majorcan Spanish. This accommodation takes place in the early stages of language contact (one of the informants had moved to the island only two and a half years before the experiment was conducted) and the main reason adduced for such an early linguistic change is the subjective factor of the individual's need of social integration in the community, a factor which reinforces the role played by the interaction among individuals in a given social network.

The issue concludes with reviews of two books which study different aspects of language contact. The first book, edited by Bert Cornillie, José Lambert and Pierre Swiggers, reviewed by Daniela Schon, includes a collection of essays: *Linguistic identities, language shift and language policy in Europe*. The second book – *New perspectives on endangered languages. Bridging gaps between sociolinguistics, documentation and language revitalization*, edited by José Antonio Flores Farfán and Fernando F. Ramallo – pays special attention to language endangerment and revitalization; it is reviewed by François Grin.

All in all, we hope that this issue will help to bring together quantitative modeling and traditional sociolinguistics and will contribute to the evolution of our understanding of the factors that operate in language contact and which may cause language attrition, language death or linguistic accommodation.

Acknowledgement

For generous financial support Lucía Loureiro-Porto is grateful to the European Regional Development Fund and the Spanish Ministry for Science and Innovation (grant FFI2011-26693-C02-02).

References

Abrams, Daniel M. & Steven H. Strogatz. 2003. Modeling the dynamics of language death. *Nature* 424. 900.

Axelrod, Robert. 2006. Agent-based modeling as a bridge between disciplines. In Leigh Tesfatsion & Kenneth L. Judd (eds.), *Handbook of computational economics*, Vol. 2: *Agent-based computational economics*, 1565–1584. Amsterdam: North Holland/Elsevier.

Baxter, Gareth J., Richard A. Blythe, William Croft & Alan J. McKane. 2009. Modeling language change: an evaluation of Trudgill's theory of the emergence of New Zealand English. *Language Variation and Change* 21. 157–196.

Benítez-Burraco, Antonio, Víctor M. Longa, Guillermo Lorenzo & Juan Uriagereka. 2008. Also sprach Neanderthalis . . . or did she? *Biolinguistics* 2(2/3). 225–232.

Benz, Anton, Christian Ebert, Gerhard Jäger & Robert van Rooij (eds.). 2011. *Language, games and evolution*. Berlin, Heidelberg & New York: Springer.

Blythe, Richard & William Croft. 2010. Can a science-humanities collaboration be successful? *Adaptive Behavior* 18. 12–20.

Bornkessel-Schlesewsky, Ina & Matthias Schlesewsky. 2009. *Processing syntax and morphology: a neurocognitive perspective*. Oxford: Oxford University Press.

Buch, Armin, David Erschler, Gerhard Jäger & Andrei Lupas. 2011. Towards automated language classification: a clustering approach. In *Proceedings of the workshop on comparing approaches to measuring linguistic differences*. http://www.sfs.uni-tuebingen.de/~gjaeger/cgi/publications.shtml (accessed 15 November 2011).

Castelló, Xavier, Andrea Baronchelli & Vittorio Loreto. 2009. Consensus and ordering in language dynamics. *European Physical Journal B* 71. 557–564.

Castelló, Xavier, Riitta Toivonen, Víctor M. Eguíluz, Jari Saramäki, Kimmo Kaski & Maxi San Miguel. 2007. Anomalous lifetime distributions and topological traps in ordering dynamics. *Europhysics Letters* 79. 66006-1–66006-6.

Castelló, Xavier, Riitta Toivonen, Víctor M. Eguíluz, Lucía Loureiro-Porto, Jari Sarmäki, Kimmo Kaski & Maxi San Miguel. 2008. Modelling language competition: bilingualism and complex social networks. In Andrew D. M. Smith, Kenny Smith & Ramon Ferrer i Cancho (eds.), *The evolution of language. Proceedings of the 7th International Conference (EVOLANG7)*, 59–66. Singapore: World Scientific Publishing.

Chapel, Laetitia, Xavier Castelló, Claire Bernard, Guillaume Deffuant, Víctor M. Eguíluz, Sophie Martin & Maxi San Miguel. 2010. Viability and resilience of languages in competition. *PLoS One* 5(1). e8681.

Crystal, David. 2000. *Language death*. Cambridge: Cambridge University Press.

de Oliveira, Paulo Murilo Castro, Søren Wichmann, Dieter Stauffer & Suzana Moss de Oliveira. 2008. A computer simulation of language families. *Journal of Linguistics* 44. 659–675.

Gong, Tao, James W. Minett & William S.-Y. Wang. 2008. Exploring social structure effect on language evolution based on a computational model. *Connection Science* 20. 135–153.
Jäger, Gerhard. 2008. Game-theoretical pragmatics. In Johan van Benthem & Alice ter Meulen (eds.), *Handbook of logic and language*, 2nd edn., 467–491. Amsterdam et al.: Elsevier.
Kandler, Anne. 2009. Demography and language competition. *Human Biology* 81 (2/3). 181–210. http://digitalcommons.wayne.edu/humbiol/vol81/iss2/5 (accessed 15 November 2011).
Köhler, Reinhard. 2012. *Quantitative syntax analysis*. Berlin & New York: Mouton de Gruyter.
Labov, William. 1994. *Principles of linguistic change. Internal factors*. Oxford: Blackwell.
Lieberman, Erez, Jean-Baptiste Michel, Joe Jackson, Tina Tang & Martin A Nowak. 2007. Quantifying the evolutionary dynamics of language. *Nature* 449. 714–716.
Liu, Haitao. 2008. The complexity of Chinese syntactic dependency networks. *Physica A* 387. 3048–3058.
Loreto, Vittorio, Andrea Baronchelli & Luc Steels. 2008. In-depth analysis of the Naming Game dynamics: the homogeneous mixing case. *International Journal of Modern Physics C* 19(5). 785–812.
Minett, James W. & William S-Y. Wang. 2008. Modelling endangered languages: the effects of bilingualism and social structure. *Lingua* 118. 19–45.
Mira, Jorge, Luis F. Seoane & Juan J. Nieto. 2011. Importance of interlinguistic similarity and stable bilingualism when two languages compete. *New Journal of Physics* 13. http://iopscience.iop.org/1367-2630/13/3/033007 (accessed 15 November 2011).
Nerbonne, John. 2010. Measuring the diffusion of linguistic change. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365. 3821–3828.
Patriarca, Marco & Els Heinsalu. 2009. Influence of geography on language competition. *Physica A* 388. 174–186.
Petroni, Filippo & Maurizio Serva. 2010. Measures of lexical distance between languages. *Physica A: Statistical Mechanics and its Applications* 389(11). 2280–2283.
Swadesh Morris. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21. 121–137.
Tamariz, Monica, Tao Gong & Gerhard Jäger. 2011. Investigating the effect of prestige on the diffusion of linguistic variants. In Laura Carlson, Christoph Hoelscher & Thomas F. Shipley (eds.), *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*, 1491–1496. Austin: Cognitive Science Society.
Wichmann, Søren & Eric W. Holman. 2008. Population size and rates of language change. Presented at the conference, *Demographic Processes and Cultural Change*, AHRC Centre for the Evolution of Cultural Diversity, University College London, 2–6 September.
Wichmann, Søren, Eric W. Holman, Dik Bakker & Cecil H. Brown. 2010. Evaluating linguistic distance measures. *Physica A: Statistical Mechanics and its Applications* 389(17). 3632–3638.
Wickstrom, Bengt-Arne. 2005. Can bilingualism be dynamically stable? A simple model of language choice. *Rationality and Society* 17(1). 81–115.
Zanette, Damian H. 2008. Demographic growth and the distribution of language sizes. *International Journal of Modern Physics C* 19. 237–247.
Zhou, Shuigeng, Guobiao Hu, Zhongzhi Zhang & Jihong Guan. 2008. An empirical study of Chinese language networks. *Physica A* 387. 3039–3047.