FALE
FACULDADE
DE LETRAS

UFMG

Computational modeling and simulation in usage-based linguistics: a necessary approach

Dalmo Buzato & Evandro L. T. P. Cunha
Faculty of Letters
Universidade Federal de Minas Gerais
Brasil

Computational linguistics and complexity sciences: a common beginning

Warren Weaver (1894 – 1978)

Using computers to translate documents between natural languages, letter to Norbert Wiener in **1947**

He published the essay "*Science and Complexity*" in **1948**

Computational linguistics and complexity sciences: a common beginning

Warren Weaver (1894 – 1978)

Using computers to translate documents between natural languages, letter to Norbert Wiener in **1947**

He published the essay "*Science and Complexity*" in **1948**

- simplicity
- disorganized complexity
- organized complexity

Computational linguistics and complexity sciences: a common beginning

Warren Weaver (1894 – 1978)

Using computers to translate documents between natural languages, letter to Norbert Wiener in **1947**

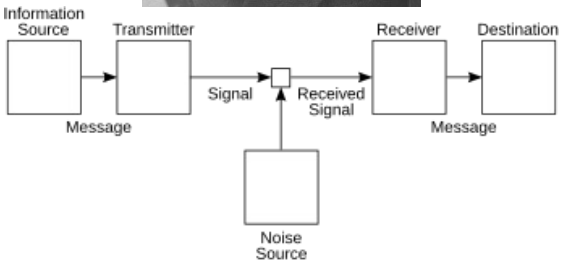
He published the essay "*Science and Complexity*" in **1948**

"Problems which involve dealing simultaneously with a sizable number of factors which are interrelated into a organic whole."

organized complexity

Computational linguistics and complexity sciences: a common beginning

Warren Weaver (1894 – 1978)



Using computers to translate documents between natural languages, letter to Norbert Wiener in **1947**

He published the essay "*Science and Complexity*" in **1948**

He published with Claude E. Shannon the essay "*The Mathematical Theory of Communication*" in **1948**

Warren Weaver (1894 – 1978)

Using computers to translate documents between natural languages, letter to Norbert Wiener in **1947**

He published the essay "*Science and Complexity*" in **1948**

He published with Claude E. Shannon the essay "*The Mathematical Theory of Communication*" in **1948**

In July **1949** he published the memorandum "*Translation*", pioneering the first methods for machine translation

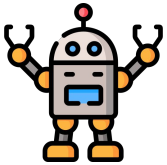
NLP/CL

Rule-based models

Example-based models

Supervised learning-based models

Transformers, BERT, LLM models



CS/Complex Systems

Information theory

Game theory

Network science
Graph theory

Artificial life

Economy

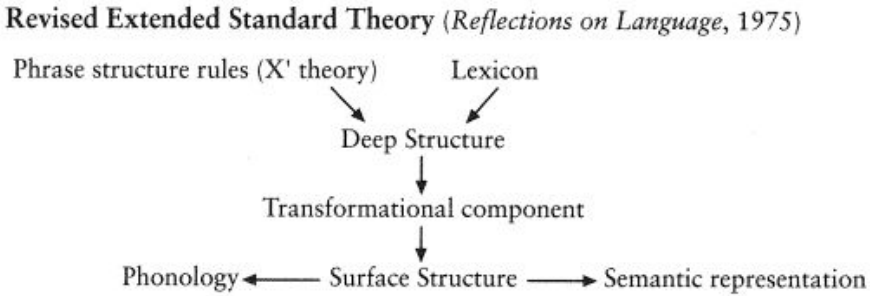
Epidemiology

Sociology

What is the space of complex systems in linguistics?

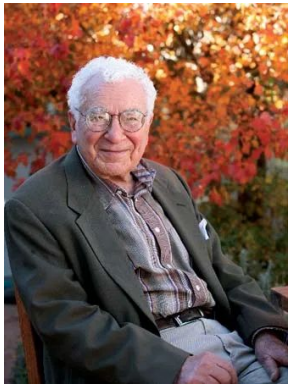
Complex systems: "A system composed of many parts that interact with each other in a ***non-linear*** and ***decentralized*** way."

Generative linguistics paradigm

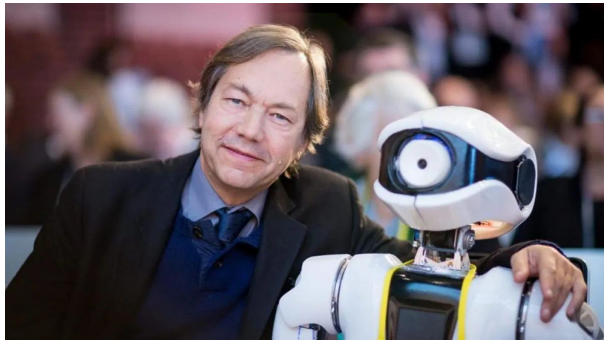


"Phonological and semantic structures are outputs of a syntactic derivation, with no significant generative capacities of their own." Jackendoff (2003, p. 196)

What is the space of complex systems in linguistics?

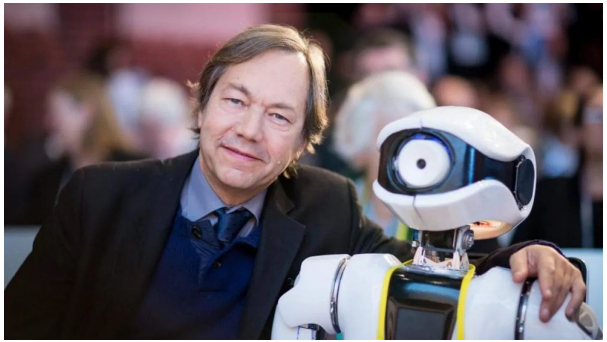


Murray **Gell-Mann** (physicist)
Sergei **Starostin** (linguist)
Merritt **Ruhlen** (linguist)
Santa Fe Institute (USA)
Language universals and historical linguistics



Luc **Steels** (computational linguist)
Bart **de Boer** (linguist)
Vrije Universiteit Brussel (Belgium)
Language evolution
Self-organization in phonological systems

What is the space of complex systems in linguistics?



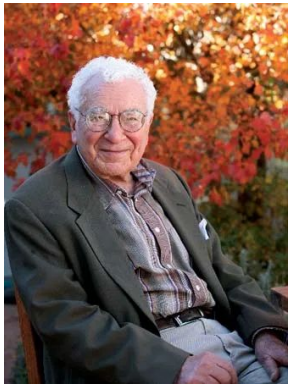
Luc **Steels** (computational linguist)
Bart **de Boer** (linguist)
Vrije Universiteit Brussel (Belgium)
Language evolution
Self-organization in phonological systems

Steels (1997): "The synthetic modeling of language origins"

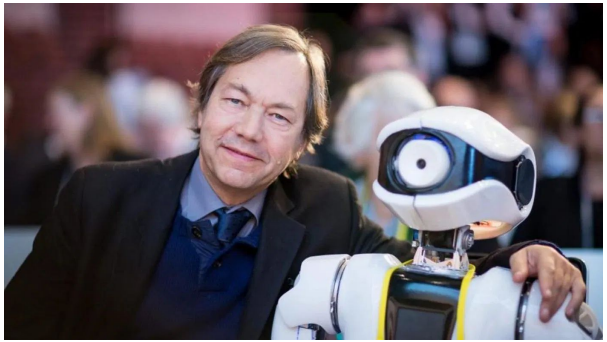
Dik, S. (1980) Studies in Functional Grammar.
Langacker, R.W. (1986) Foundations of Cognitive Grammar.
McClelland, J.L., and Rumelhart, D.E. eds. (1986). Explorations in Parallel Distributed Processing.
Thomason, S.G., T. Kaufman (1988) Language Contact, Creolization, and Genetic Linguistics.

Steels, L. (2017). Basics of fluid construction grammar. *Constructions and frames*, 9(2), 178-225.

What is the space of complex systems in linguistics?



Murray **Gell-Mann** (physicist)
Sergei **Starostin** (linguist)
Merritt **Ruhlen** (linguist)
Santa Fe Institute (USA)
Language universals and historical linguistics



Luc **Steels** (computational linguist)
Bart **de Boer** (linguist)
Vrije Universiteit Brussel (Belgium)
Language evolution
Self-organization in phonological systems

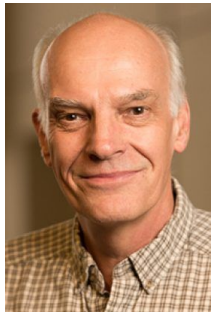
William S-Y. **Wang** (linguist)
Tao **Gong** (linguist)
Hong Kong Polytechnic University
Lexical diffusion and language change

What is the space of complex systems in linguistics?

"*Continued Study of Language Acquisition and Evolution*" working group meeting, Santa Fe Institute, March 1–3, 2007.

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D., & Schoenemann, T. (2009). **Language Is a Complex Adaptive System: Position Paper**. Language Learning, 59, 1–26.




Joan Bybee, Morten Christiansen, William Croft, Nick Ellis, Diane Larsen-Freeman

What is the space of complex systems in linguistics?

Modeling Usage-Based Acquisition and Change

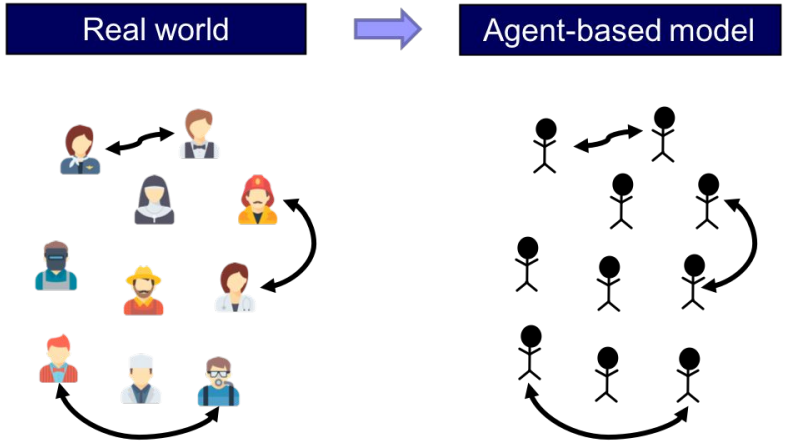
In the various aspects of language considered here, it is always the case that form, user, and use are inextricably linked. However, such complex interactions are difficult to investigate *in vivo*. Detailed, dense longitudinal studies of language use and acquisition are rare enough for single individuals over a time course of months. Extending the scope to cover the community of language users, and the timescale to that for language evolution and change, is clearly not feasible. Thus, our corpus studies and psycholinguistic investigations try to sample and focus on times of most change and interactions of most significance. However, there are other ways to investigate how language might emerge and evolve as a CAS. A valuable tool featuring strongly in our methodology is mathematical or computational modeling.

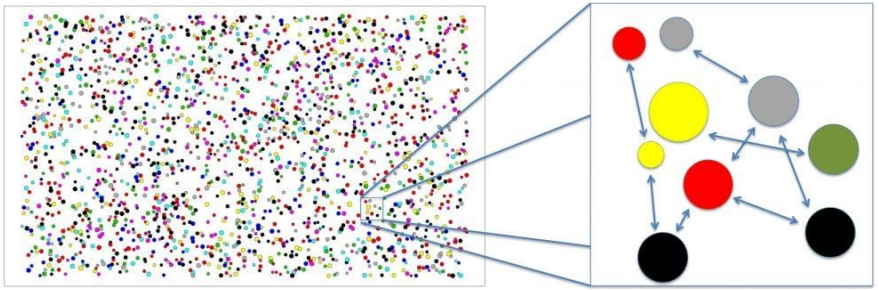
Beckner et al. (2009)

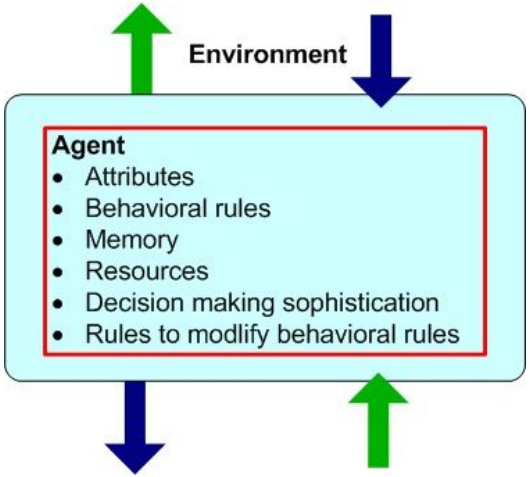
Concrete language use
vs.
Modeled/simulated language use

It enables investigations limited by experimental or corpora studies

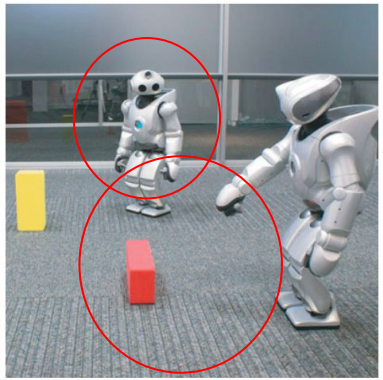
Agent-based modeling (ABM)

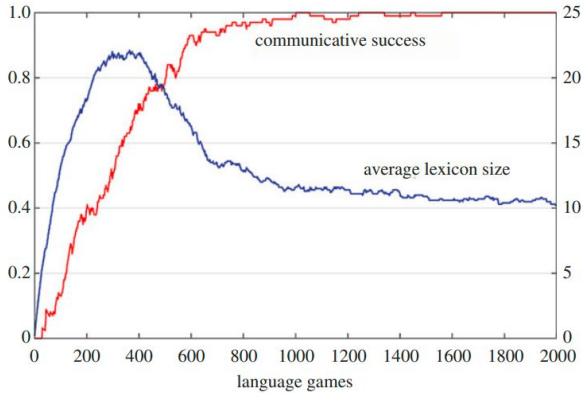


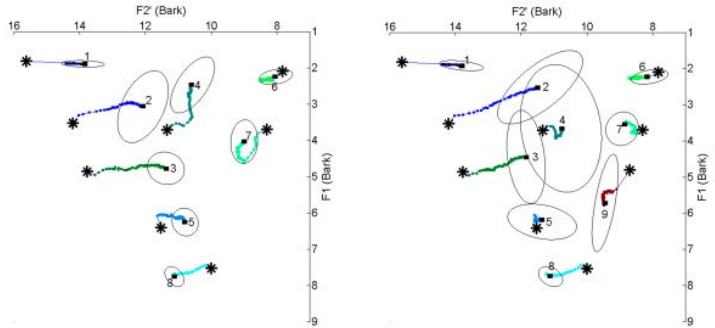




Research examples







(a) Eight-Vowel Simulations (b) Nine-Vowel Simulations

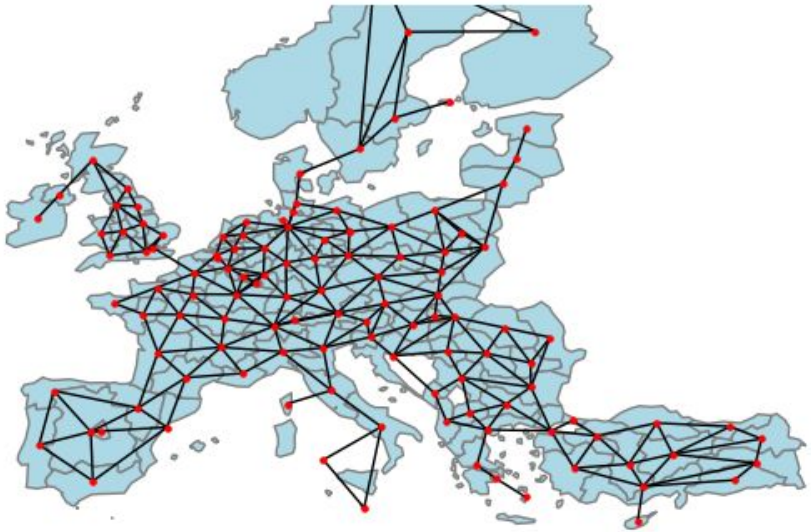
Agent-based study for the origins of words (Steels, 2016)

Ostensive-inferential communication (Sperber and Wilson, 1986)

Contact-induced language change in Xumi (旭米), an unwritten Tibeto-Burman language (Chirkova and Gong, 2014)

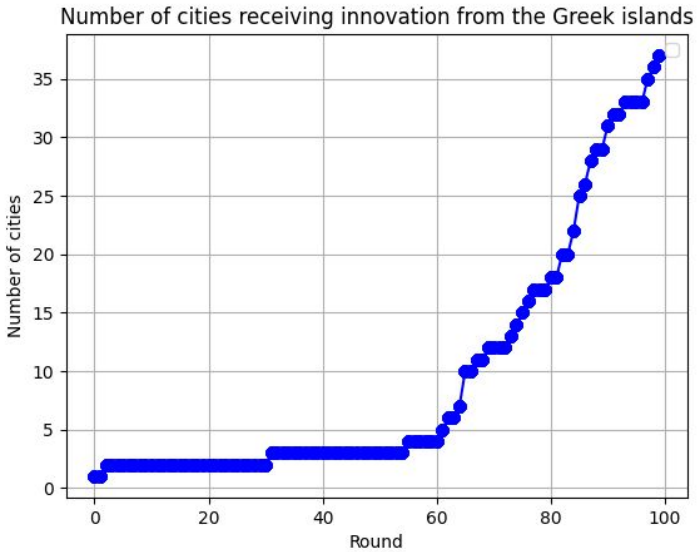
Research examples

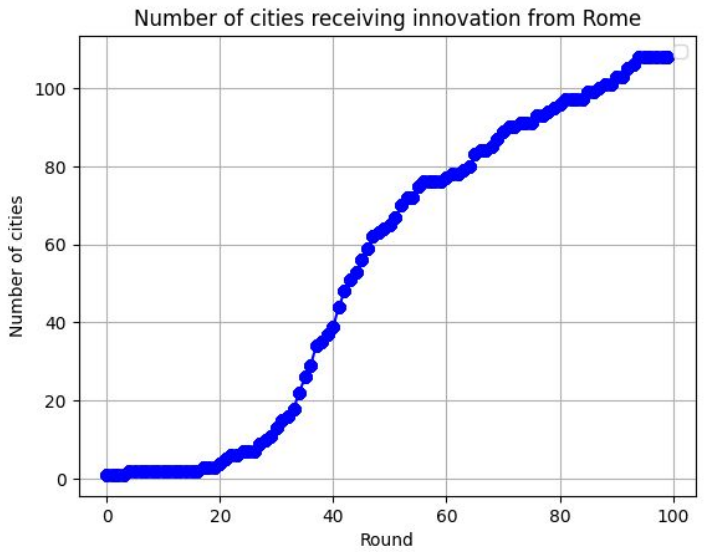
Dalmo Buzato and Evandro L. T. P. Cunha. 2024. Bartoli's areal norms revisited: an agent-based modeling approach. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese (PROPOR)*, pages 422– 431.



Testing Bartoli's norms about **language change** and the **geographical space** (center, periphery and isolated area).

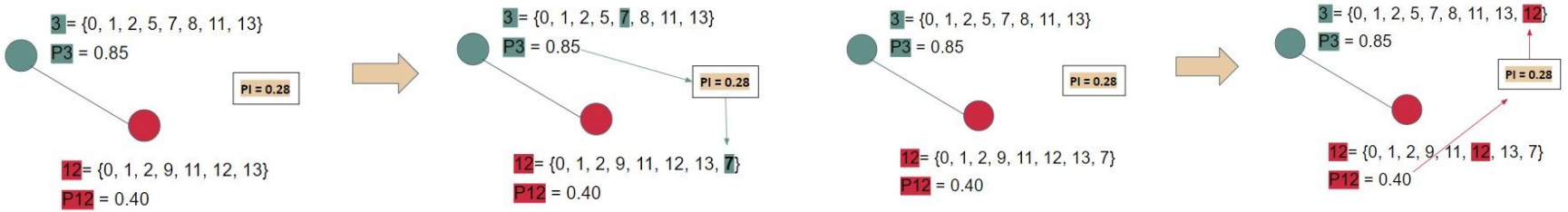
Bartoli's theory is based on examples and counterexamples from Romance linguistics but without quantitative validation.


Number of cities receiving innovation from the Greek islands
Number of cities
Round
0
5
10
15
20
25
30
35
0
20
40
60
80
100


Number of cities receiving innovation from Rome
Number of cities
Round
0
20
40
60
80
100
0
20
40
60
80
100

Research examples

Dalmo Buzato and Evandro L. T. P. Cunha. 2024. Agent-based modeling of language change in a small-world network. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 594–599.

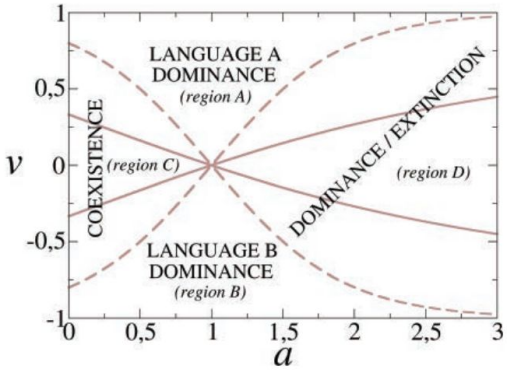


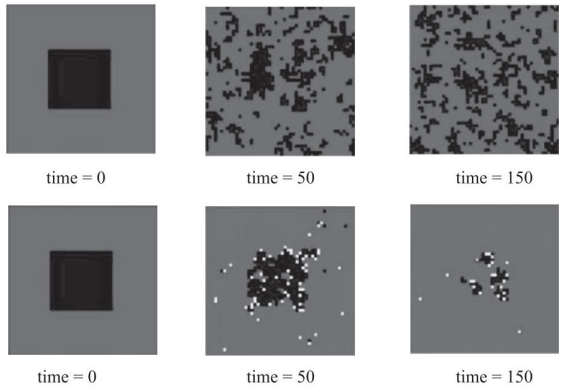
The notion of prestige was inspired by Castelló, Loureiro-Porto and San Miguel (2013) and from traditional variationist sociolinguistics.

The change occurs not at the level of languages (choice between language A or B), but at the level of idiolect (propagation of items).

Brazilian reception

Viotti (2020): "[...] *ethnographies seem to more adequately capture the dynamicity and complexity of language as a social phenomenon than some computational models that need to greatly simplify the characterization of the system, and that are based on choices made by researchers rather than on factors that, in fact, are defining of the system.*"





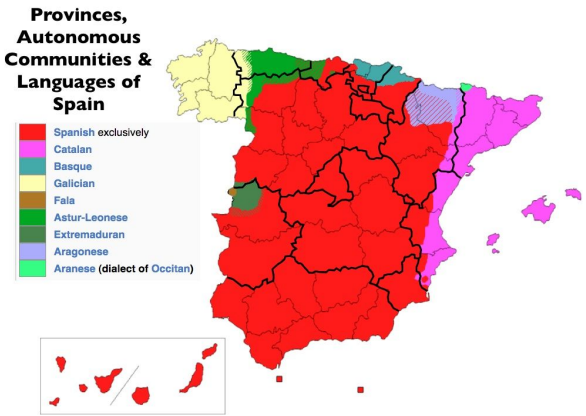
Castelló, Loureiro-Porto and San Miguel (2013)

Sociolinguistic typology (Croft, 2021)

Language contact and competition situations:

Castelló, Loureiro-Porto and San Miguel (2013): Contact, competition and language shift in Spain and the Mediterranean

Viotti (2020): Small-scale multilingualism in Amazonia (Lüpke, 2016; Lüpke et al., 2020; Croft, 2021)



Ongoing research (2024 and 2025)

"All models are wrong, but some are useful." George E. P. Box

Southern Amazon (Rondônia)

Impressive language diversity (Van der Voort, 2023 and Lüpke et al., 2020)

17 languages; **8** language families; there is the presence of isolated languages

Despite linguistic differences, atypical cultural similarities

Maldi (1991): Complexo Cultural do Marico

Lista comparativa de línguas

Português	Arikapú	Jabuti	Makurap	Ajuru	Koaratira/ Sakirap	Aruá	Tupari
água	bi	bzürü	ü	ügü	ükü	ü	ü
fogo	pikô	pitié	uaxát	aokap	utat	káin	kupkap
milho	titi	titi	atiti	atiti	atiti	maék	pupáp
macaxeira	boré	boré	manü	manü	tapcit	pabüiá	máin
homem	uananhé	tüê	kitô	baikop	mankup	woi	ukin
mulher	pakué	pakô	arampinhã	araminá	araminá	uazenp	araminá
civilizado	eré	eré	eré	uerep	guerep	goián	talipá
peixe	minon	minon	putkap	iboi	küpit	borip	ipot
onça	kurá	uá	amekô	amekô	amekô	nenkó	amekô
sol	tahan	tohon	gueát	jacop	tuakop	ngát	kiakop
lua	kupá	kupá	ulí	pakuri	pakuri	gatí	kuepá

Tupi-Tupari

Tupi-Mondé (Aruá)

Tupi-Tupari (Koaratira/Sakirap)

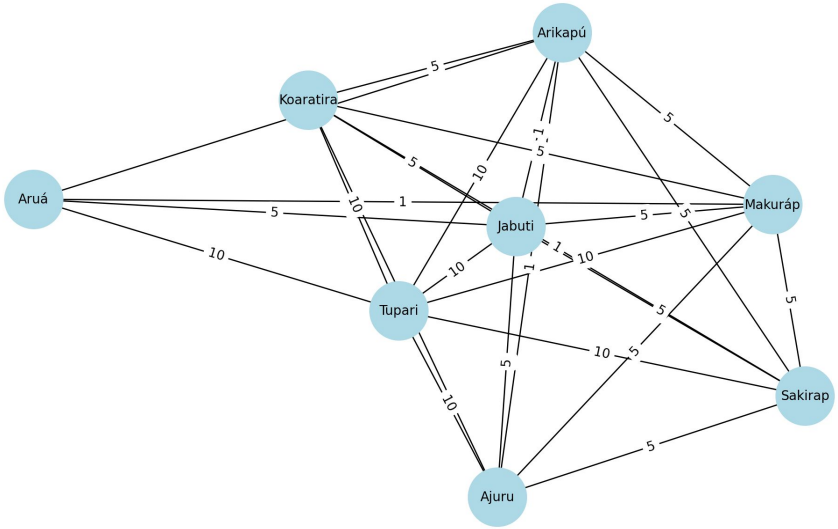
Jabuti

SOCIEDADES DOS RIOS BRANCO, COLORADO E MEQUENS

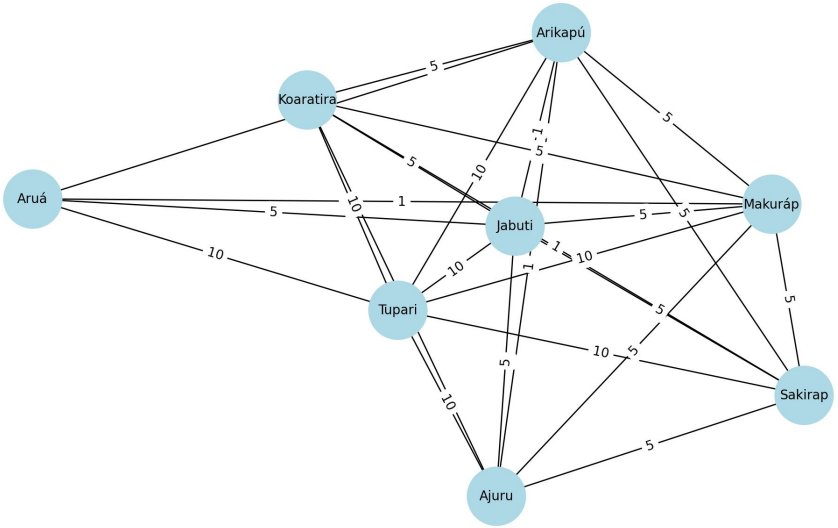
SOCIEDADE	LOCALIZAÇÃO TRADICIONAL	PROXIMIDADE		
		PERTO	LONGE	MUITO LONGE
Jabuti	acima das cabeceiras do rio Branco. Aldeias na margem esquerda do curso alto	Arikapú	Ajuru Aruá Makurap	Tupari
Arikapú	margem esquerda do alto rio Branco, território contíguo à área Jabuti, mas mais abaixo	Jabuti Ajuru	Aruá Makurap	Tupari
Ajuru	entre as cabeceiras do rio Colorado e as cabeceiras do rio Terebito	Arikapú	Jabuti Makurap	Tupari
Makurap	entre as cabeceiras do rio Branco, mais afastados da margem esquerda e ambas as margens do alto rio Colorado	Aruá	Arikapú Jabuti	Tupari
Aruá	igarapé Gregório, alto rio Branco	Makurap	Arikapú Jabuti	Tupari
Koaratira	alto rio Mequens	Sakirap	Makurap Ajuru Arikapú Jabuti	Tupari
Sakirap	alto rio Verde, afluente do Corumbiara	Koaratira	Makurap Ajuru Arikapú Jabuti	Tupari

SOCIEDADES DOS RIOS BRANCO, COLORADO E MEQUENS

SOCIEDADE	LOCALIZAÇÃO TRADICIONAL	PROXIMIDADE		
		PERTO	LONGE	MUITO LONGE
Jabuti	acima das cabeceiras do rio Branco. Aldeias na margem esquerda do curso alto	Arikapú	Ajuru Aruá Makurap	Tupari
Arikapú	margem esquerda do alto rio Branco, território contíguo à área Jabuti, mas mais abaixo	Jabuti Ajuru	Aruá Makurap	Tupari
Ajuru	entre as cabeceiras do rio Colorado e as cabeceiras do rio Terebito	Arikapú	Jabuti Makurap	Tupari
Makurap	entre as cabeceiras do rio Branco, mais afastados da margem esquerda e ambas as margens do alto rio Colorado	Aruá	Arikapú Jabuti	Tupari
Aruá	igarapé Gregório, alto rio Branco	Makurap	Arikapú Jabuti	Tupari
Koaratira	alto rio Mequens	Sakirap	Makurap Ajuru Arikapú Jabuti	Tupari
Sakirap	alto rio Verde, afluente do Corumbiara	Koaratira	Makurap Ajuru Arikapú Jabuti	Tupari



Next steps...



Calculate similarity indices and distance matrices between languages using wordlists.

Use epidemiological models to analyze the spread of items between groups.

Think about how to include the Kanoé (isolated group identified in the 80s, first linguistic description only in 2001).

Do your models serve us?

Viotti's argument is partially valid for us: the models created by Spanish colleagues are a great advance, but **they do not explain sociolinguistic relations in the Amazonia**.

There is probably a **strong relationship** between **language contact** and the **language evolution**, and small-scale multilingualism could help us understand the grammatical and social evolution of human language.

Thinking about models that **overcome prestige** as a condition for language change is thinking about valid models for Amazonian languages, but also models that reveal aspects about the origin of language.

Summing up...

Computational modeling and simulations in linguistics are possible within the usage-based framework.

These methodologies do not **replace**, but rather **complement**, data from psycholinguistics or corpus linguistics.

It is possible to model many things, but you need the right tools (programming, statistics, mathematical notions).

Language is a complex adaptive system (and this view unites language with several other systems in the universe!)

References

JACKENDOFF, Ray. Précis of foundations of language: Brain, meaning, grammar, evolution. Behavioral and brain sciences, v. 26, n. 6, p. 651-665, 2003.

STEELS, Luc. Basics of fluid construction grammar. Constructions and frames, v. 9, n. 2, p. 178-225, 2017.

BECKNER, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D., & Schoenemann, T. (2009). Language Is a Complex Adaptive System: Position Paper. Language Learning, 59, 1–26.

STEELS, Luc. Agent-based models for the emergence and evolution of grammar. Philosophical Transactions of the Royal Society B: Biological Sciences, v. 371, n. 1701, p. 20150447, 2016.

SPERBER, D. and Wilson, D. (1986) Relevance: Communication and Cognition, Blackwell.

CHIRKOVA, Katia; GONG, Tao. Simulating vowel chain shift in Xumi. Lingua, v. 152, p. 65-80, 2014.

BUZATO, Dalmo and Evandro L. T. P. Cunha. 2024. Bartoli's areal norms revisited: an agent-based modeling approach. In Proceedings of the 16th International Conference on Computational Processing of Portuguese (PROPOR), pages 422– 431.

BUZATO, Dalmo and Evandro L. T. P. Cunha. 2024. Agent-based modeling of language change in a small-world network. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 594–599.

CASTELLÓ, Xavier; LOUREIRO-PORTO, Lucía; SAN MIGUEL, Maxi. Agent-based models of language competition. International journal of the sociology of language, v. 2013, n. 221, p. 21-51, 2013.

VIOTTI, Evani. Avaliando a vitalidade linguística em contextos de multilinguismo: etnografias versus modelos computacionais. Revista Linguíjtica, v. 16, n. 1, p. 62-84, 2020.

MALDI, Denise. O complexo cultural do Marico: sociedades indígenas dos rios Branco, Colorado e Mequens, afluentes do Médio Guaporé. Boletim do Museu Paraense Emílio Goeldi, v. 7, n. 2, p. 209-269, 1991.

FALE
FACULDADE
DE LETRAS

UFMG

Muito obrigado! Thank you!

dalmobuzato@ufmg.br
cunhae@ufmg.br

VII Simpósio Internacional de Linguística Funcional (SILF)

Belo Horizonte, Minas Gerais

September 4-6, 2024