Identifying probable pathways of language diffusion in South America

Peter Ranacher
University of Zurich,
Dept. of Geography
Winterthurerstrasse 190
Zurich, Switzerland
peter.ranacher@geo.uzh.ch

Rik van Gijn
University of Zurich, Dept.
of Comparative Linguistics
Plattenstrasse 54
Zurich, Switzerland
erik.vangijn@uzh.ch

Curdin Derungs
University of Zurich,
URPP Language and Space
Freiestrasse 16
Zurich, Switzerland
curdin.derungs@geo.uzh.ch

Abstract

In linguistics, there is broad consensus that river networks have played an important role in the diffusion of languages in South America. However, the presence of a river alone does not imply language diffusion. Languages have spread along specific routes in the network. It is largely unknown where these routes are located, since evidence of language diffusion is often sparse and only spatially-implicit. In this paper we propose an approach to identify probable pathways of language diffusion along the Amazon River network, combining ideas from route planning and route inference. Route planning proposes possible routes of linguistic diffusion along the river network. Route inference tests these against evidence from linguistic data. We find significant evidence for language diffusion along few, specific branches of the Amazon. Our approach is not restricted to linguistic data alone. It is generally suitable to explore deep-time processes in space for which evidence is sparse and spatially-implicit.
Keywords: route simulation, route inference, Amazon River, language contact, deep-time process

1 Introduction

Imagine you know proximate locations and detailed linguistic characteristics of some 100 languages, distributed over the Amazon region, an area three times the size of Germany. How would you tackle the hypothesis that some of these languages are subject to a common diffusion process in space and time? How would you identify probable pathways of diffusion? In this study we contribute to this question by linking route planning heuristics to route inference, two well researched topics in GIScience.

Route planning is concerned with finding suitable paths in a network (Hofmann-Wellenhof, et al., 2011). A suitable path minimizes costs – e.g. in terms of travel time, distance or money – and maximizes utility – e.g. most scenic route (Runge, et al., 2016) or optimal route for disabled people (Neis & Zielstra, 2014). A classic example of route planning is the travelling salesman problem. A salesperson must deliver goods to customers in a city. What is the shortest route to dispatch the goods?

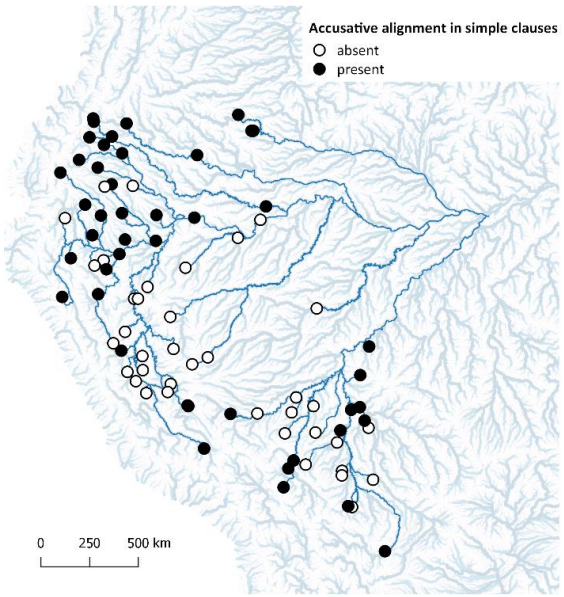
Route inference is concerned with reengineering probable routes from incomplete or sparse movement data (Rahmani & Koutsopoulos, 2013). Let us assume that the above salesperson carries a logbook, where he/she records his/her route. A classical route inference task is to identify the salesperson's route from the descriptions in the logbook. If descriptions are sufficiently detailed and spatially explicit, route inference results in a definite answer on the salesperson's whereabouts in space (and time). In a network, this task is commonly known as map-matching (Rahmani & Koutsopoulos, 2013).

However, descriptions might be incomplete or spatially only implicit. In the logbook, some important entries might be missing. The available entries might conceal spatial evidence in seemingly non-spatial information (e.g. the salesperson lists the customers' names, rather than their addresses). Trying to infer clear-cut routes from such information will certainly fail. Nevertheless, some routes can still be assumed to be more probable than others, simply because they reflect the overall task of a travelling salesperson. These routes can be found with *route planning*. Of these only some are in agreement with the sparse and implicit information in the logbook. These routes can be identified with *route inference*. Thus, the combination of route planning and route inference allows to identify probable routes, given uncertain, sparse or implicit information, similar to Wei, et al. (2012). Related case studies in GIScience typically aim at reconstructing path probabilities from GPS data associated with uncertainties of some 5 to 300 meters (Tang, et al., 2016) or 5 to 10 minutes variation in travel time (Chen, et al., 2016).

In this paper, we focus on a problem similar in character, but considerably different in the available information as well as scale, both spatial and temporal: the reconstruction of language diffusion in South America. Language diffusion comprises two aspects: *language contact*, where speakers of two or more languages interact, and *language expansion*, where one language family spreads. Though different from a linguistic perspective, both yield the same results: languages often far dispersed in space having similar characteristics. For an overview on spatial analysis in historical linguistics, please refer to Haynie (2014).

The diffusion under study has covered hundreds of years and vast parts of the South American continent. In terms of route planning, we test the hypothesis that diffusion has occurred along river networks. As input information for route inference

Figure 1: Presence and absence of Acoustic alignment for simple clauses in the 84 languages along the Amazon River network



(using the analogy of the traveling salesperson's logbook) we use detailed linguistic characteristics of some 84 languages, as well as their current spatial distribution.

2 Language diffusion in the Amazon river network

The Amazon River network is the largest river system on Earth. It is home to more than 350 ethno-linguistic groups belonging to several different language families (Aikhenvald, 2012). We know that rivers played a central role as pathways of movement in the spread of language families and contact between ethnic groups in Amazonia, e.g. see Eriksen (2011). However, evidence suggests that river networks are not always good predictors for the distribution of linguistic features (van Gijn, et al., in press). This surprising result is due to several factors, including scale and human choice. The large networks surveyed in van Gijn et al. (in press) may encompass several local histories that cancel out each other's signals. In addition, the presence of a river network does not deterministically imply contact or expansion. What is needed is a more fine-grained approach that can identify favoured branches of language diffusion, i.e. parts of the river networks that are good predictors of clusters of linguistic features.

In this manuscript we introduce such an approach. We first propose possible diffusion processes (route planning). Then we test each process against evidence from linguistic data (route inference). The approach allows to identify probable pathways of linguistic contact or expansion, thus increasing our understanding of the population history of this part of South America.

2.1 Data

The Amazon River took its present shape approximately 2.4 Million years ago (Hoorn, et al., 2017), whereas South America was only populated around 15-23 thousand years ago (Nielsen, et al., 2017). We, therefore, use modern day line geometries and neglect any changes of the river network during the rather short time period relevant to our study. The Amazon River network was derived from the *HydroSHEDS*[1] data set provided by the U. S. Geological Survey. The *HydroSHEDS River Network* data comprise line geometries of rivers on a global scale at a resolution of 15 sec (Lehner, et al., 2006). By adding an explicit topology to the set of line vectors, we created a routable network of the river system.

We collected data on 23 linguistic features on the basis of written sources on the languages (Table 1). All features are known to have diffused across the Amazon and adjacent Andean regions (van Gijn, 2014). The language locations are based on the point data set provided by Hammarström et al. (2016). Using point data for languages can be justified given the generally small areas covered by individual language groups (with usually just a few hundred or few thousand speakers). In order to situate the languages on the river network, we matched each language point to the closest river branch.

Figure 1 shows all 84 languages along the river network and the feature *Accusative alignment in simple clauses*. Accusative alignment is defined as the identical behaviour of subjects of intransitive and transitive clauses, as opposed to objects of transitive clauses. Black dots indicate presence (languages that have the feature), white dots indicate absence (languages that don't have the feature).

2.2 Methodology

We propose the following approach to generate a distribution of possible diffusion scenarios. We test some 5 million individual diffusion processes, defined as a concatenation of language points. Diffusion processes start at a random node and propagate upstream. At each network confluence the diffusions randomly choose between three options:

a. follow one of the two branches,
b. split and follow both branches,
c. terminate.

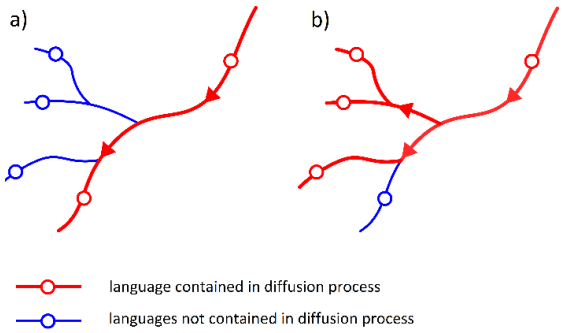
We vary the probabilities of the three options of branching. Thus, we generate diverse types of diffusion processes, ranging from broad diffusions along many branches, to narrow and target-oriented diffusions along only a few branches (see Figure 2).

Each randomly generated diffusion process can now be represented as a set of languages. In order to avoid diffusion along branches without languages, the network is pruned, such that all upstream branches lead to at least one language.

We compare the distribution of linguistic features in the set of languages to the distribution of linguistic features in all

[1] https://hydrosheds.cr.usgs.gov/

Figure 2: Target-oriented (a) and broad diffusion (b)



South American languages. To guarantee for robustness, language sets with less than four instances are discarded.

The intuition behind the statistical test is as follows: Some linguistic features are frequent, others are rare in South American languages. If a generally absent feature is frequently present in a diffusion process we are likely to observe evidence of linguistic diffusion (the same holds true if a generally present feature is frequently absent in the diffusion process).

Table 1 shows the expected probability (pr) of all 23 features to be present (absent) in South American languages.

In a random sample of size n taken from all South American languages, one expects a feature F to be present (absent) with the binomial distribution

$$X \sim B(n, pr). \tag{1}$$

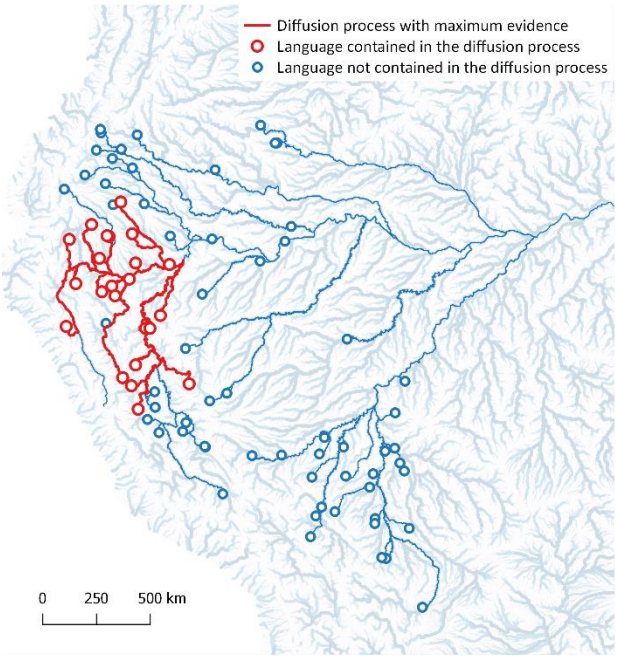
X is the null distribution of F in a sample of size n. X is calculated for each feature and each state (presence, absence) according to the data in Table 1. The null hypothesis states that the presence (absence) in the sample follows the null distribution. For each F we carry out an exact one-tailed binomial test to evaluate the frequency in the sample against the null distribution. Rejecting the null hypothesis implies that a feature is significantly more often present (absent) than expected, which we consider evidence of language diffusion. When the null hypothesis is rejected for multiple features the evidence grows.

Table 1: The features and their expected probability (pr) in South America.

No	Feature	pr (absence)	pr (presence)	Source
1	Central high vowel	0.4	0.6	1
2	Mid vowels	0.26	0.74	2
3	Contrastive vowel nasalization	0.52	0.48	3
4	Palatal nasal	0.6	0.4	1
5	Velar-uvular opposition for stops	0.92	0.08	1
6	Retroflex affricates	0.94	0.06	1
7	More affricates than fricatives	0.89	0.11	2
8	Single liquid phoneme	0.52	0.48	2
9	Simple syllable structures (no of consonants allowed in coda)	0.31	0.69	2
10	Nasal spread	0.87	0.13	2
11	Glottalized stops	0.93	0.07	1
12	Aspirated stops	0.84	0.16	1
13	Prefixes (specify whether this includes person prefixes)	0.69	0.31	3
14	Isomorphism of possessor and core verbal argument person markers	0.14	0.86	4
15	Small elaborate case marking system	0.56	0.44	3
16	Core case (i.e. case marker for nominative, accusative, ergative, absolutive)	0.57	0.43	5
17	Accusative alignment in simple clauses	0.69	0.31	3
18	Dependent marking for possession (beyond pronouns)	0.58	0.42	6
19	Classifier or gender systems	0.5	0.5	6
20	O before A basic constituent order	0.91	0.09	3
21	Basic AN order	0.75	0.25	3
22	Numerals > 9 (non-Spanish/Portuguese)	0.54	0.46	2
23	Ideophones (as a clear separate word class)	0.94	0.06	2

Source: 1: Michael et al. (2012), 2: own data, 3: Dryer & Haspelmath (2013), 4: Siewierska (1998), 5: Birchall (2014), 6: Krasnoukhova (2012)

Figure 3: Diffusion process with maximum evidence



Languages contained in the diffusion process

Achuar	Cashibo	Kokama	Shuar
Aguaruna	Chamicuro	Muniche	Taushiro
Andoa	Chayahuita	N Pastaza Q	Ucuyali-Yurua Ash
Cajamarca Q	Cholon	Panobo	Urarina
Candoshi Shapra	Huallaga Q	S Pastaza Q	Yanesha
Capanahua	Jebero	Shipibo	

Significant features ($p < 0.03$)

Feature	P = presence A = absence	p-value
2	A	0.022
3	A	0.002
4	P	0.013
6	P	0.011
12	A	0.018
13	P	0.002
14	A	6.40E-05
15	A	0.007
16	P	0.015
17	P	0.018
18	P	0.004
19	A	0.008
21	P	0.002

3 Results

We computed ~5 million diffusion processes of which ~1.5M had a minimum number of at least four languages. We carried out ~69M exact binomial tests – one test for each of the 23 features and each of the two states (present/absent) in each diffusion process. Figure 3 shows the diffusion process with maximum numbers of statistically significant linguistic features, i.e. the pathway of the Amazon River network with highest evidence of fostering linguistic diffusion.

This diffusion process includes 23 languages (this is by coincidence the same number as the linguistic features) located in the Western part of the Amazon. Thirteen of the 23 features show significant evidence of linguistic diffusion ($p < 0.03$).

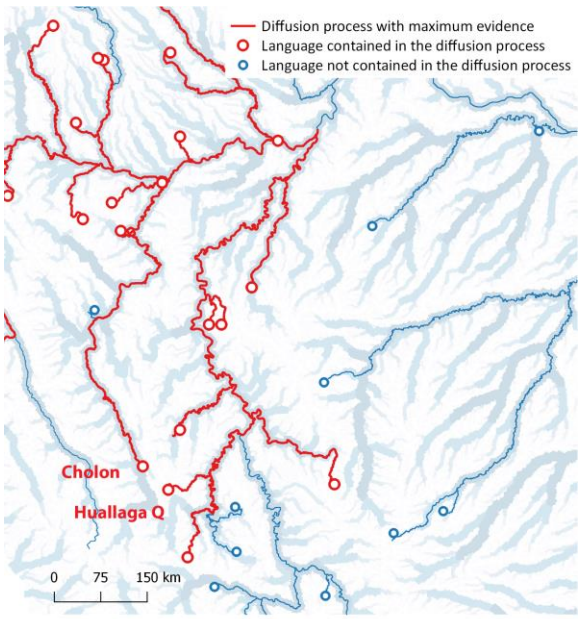
4 Discussion and future work

We started off by referring to the travelling salesman problem, its relation to route planning and route inference, and the calculation of route probabilities from incomplete data.
The state of the art approaches in GIScience often aim to compute route probabilities in an every-day context (Tang, et al., 2016). Compared to these, the setting presented in this paper is of considerably larger spatial and temporal scale, introducing a high degree of uncertainty.

Our approach allows to study diffusion processes of languages using as input present-day language locations, associated with linguistic characteristics. In particular we test the hypothesis if South American languages spread along the Amazon River network. We would like to emphasize that, despite the particularity of this case study, comparable research settings and analysis needs are commonplace in the broader humanities, ranging from anthropology to cultural or historical sciences, hence offering the potential for the above approach to be applied more widely to other, similar contexts.

We identified one diffusion process with maximum evidence (13 out of 23 features are significant; Figure 3). However, there are other diffusion processes with only slightly less significant features, i.e. 30 processes with twelve significant features and ~430 with eleven significant features. Most of these occur

Figure 4: Cholon and Huallaga Q – close in Euclidean space, far away along the river



along similar pathways of the river network and comprise similar sets of languages. Hence, there is cumulative evidence of language diffusion in a very particular part of the Amazon, evidence of a process that could be described as Inca influence on lowland languages.

The hypothesis of diffusion occurring along rivers seems to be valid for some parts of the observed realities, but not for others. Figure 4 shows an enlarged part from Figure 3. The two languages Cholon and Huallaga Q are close in space, but far away in the river network. It is more probable that language diffusion took the direct path than the detour along the river. In situations like this, one needs to be careful not to mix correlation with causality. Therefore, a linguistically and geographically informed interpretation of the results is crucial.

In this paper we discussed only one possible scenario of language diffusion, in which languages spread upstream along a river network. Other scenarios provide interesting directions for future research, for example, diffusion processes in unconstrained geographical space or along trade networks. Moreover, we assumed that all features and all languages are equally relevant in the diffusion process. Future work will focus on identifying the diffusion of different types of features (e.g. features that are known to have an Andean origin) and different types of language formation processes (e.g. spread of language families, as opposed to language contact). Another interesting topic for future research is that of optimization. The parameters of diffusion processes with significant evidence (i.e. the starting point and the branching behaviour) can be used to optimize the proposal of new diffusions.

Acknowledgements

The research reported here was supported by Grant No. CRSII1_160739 (“Linguistic morphology in time and space”) from the Swiss National Science Foundation.

References

Aikhenvald, A. Y., 2012. *Languages of the Amazon.* Oxford: Oxford University Press.

Birchall, J., 2014. *Argument marking patterns in South American languages.* Utrecht: Netherlands Graduate School of Linguistics.

Chen, B. Y. et al., 2016. Measuring place-based accessibility under travel time uncertainty. *International Journal of Geographical Information Science*, pp. 1-22.

Dryer, M. S. & Haspelmath, M. eds., 2013. *World atlas of language structures online.* Leipzig: Max Planck Institute for Evolutionary Anthropology.

Eriksen, L., 2011. *Nature and Culture in Prehistoric Amazonia Using GIS to reconstruct ancient ethnogenetic processes from archaeology, linguistics, geography, and ethnohistory.* Lund: Human Ecology Division, Lund University.

Hammarström, H., Forkel, R., Haspelmath, M. & Bank, S., 2016. Glottolog 2.7. Jena: Max Planck Institute for the Science of Human History. *Online: http://glottolog. org.*

Hofmann-Wellenhof, B., Legat, K. & Wieser, M., 2011. *Navigation: principles of positioning and guidance.* Vienna: Springer Science & Business Media.

Krasnoukhova, O., 2012. *The noun phrase in the languages of South America.* Utrecht: Netherlands Graduate School of Linguistics.

Lehner, B., Verdin, K. & Jarvis, A., 2006. HydroSHEDS technical documentation. *World Wildlife Fund US, Washington, DC*, pp. 1-27.

Michael, L., Stark, T. & Chang, W., 2012. *South American Phonological Inventory Database v1. 1.3*, Berkeley: University of California.

Neis, P. & Zielstra, D., 2014. Generation of a tailored routing network for disabled people based on collaboratively collected geodata. *Applied Geography*, Volume 47, pp. 70-77.

Rahmani, M. & Koutsopoulos, H. N., 2013. Path inference from sparse floating car data for urban networks. *Transportation Research Part C: Emerging Technologies*, Volume 30, pp. 41-54.

Runge, N., Samsonov, P., Degraen, D. & Schöning, J., 2016. *No more autobahn!: Scenic route generation using googles street view.* Sonoma, Proceedings of the 21st International Conference on Intelligent User Interfaces, pp. 147-151.

Siewierska, A., 1998. On nominal and verbal person marking. *Linguistic Typology*, Volume 2, pp. 1-55.

Tang, J., Song, Y., Miller, H. J. & Zhou, X., 2016. Estimating the most likely space--time paths, dwell times and path uncertainties from vehicle trajectory data: A time geographic method. *Transportation Research Part C: Emerging Technologies*, Volume 66, pp. 176-194.

van Gijn, R., 2014. The Andean Foothills and Adjacent Amazonian Fringe. *The Native Languages of South America: Origins, Development, Typology*, p. 102.

van Gijn, R. et al., in press. Linguistic areas, linguistic convergence, and river systems in South America. In: R. Hickey, ed. *Handbook of Linguistic Areas.* Cambridge: Cambridge University Press, pp. 964 - 996.

Wei, L.-Y., Zheng, Y. & Peng, W.-C., 2012. *Constructing popular routes from uncertain trajectories.* Beijing, Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 195-203.