Pieter Muysken, Harald Hammarström, Joshua Birchall,
Rik van Gijn, Olga Krasnoukhova, and Neele Müller

5 Linguistic areas, bottom-up or top-down? The case of the Guaporé-Mamoré[1]

1 Introduction

In the most extreme case, there are two opposite ways of looking at the notion of linguistic area or Sprachbund:

BOTTOM-UP. One starts looking for similarities between the languages of a particular region, similarities which cannot be explained through common ancestry or through typological pressure or chance. Historical linguists may thus stumble upon the "unusual" spread or patterning of features in a particular area. Once enough of these features have been found, a plausible case for a linguistic area can be made. This is a traditional way of defining a linguistic area. In this paper we will call this standard approach the bottom-up perspective, and exemplify it with the wide-ranging study of Crevels and Van der Voort (2008) on the Guaporé-Mamoré at the border of Bolivia and the Brazilian state of Rondonia.

TOP-DOWN. The alternative is that one takes a preselected, independent, and supposedly cross-linguistically valid set of features, and then tries to determine whether the distribution of these features shows areal bias, in terms of the geographic density of particular feature specifications: they are more likely to occur inside that area than outside of it.

Now that increasingly large systematic data sets are becoming available for many languages and regions and 'areal' explanations are sought after for different phenomena, it is useful to compare these approaches and test their strengths and weaknesses. Do these approaches yield similar results, or are

1 This paper was written with the support of the European Research Council "Traces of Contact" Grant to the LinC group at the Centre for Language Studies, Radboud University Nijmegen. Muysken is responsible for the overall concept and the final text. Birchall coded the argument marking features. Van Gijn coded the subordination features. Hammarström developed the quantitative models for testing sprachbunds. Krasnoukhova coded the Noun Phrase features, and Müller the TAME features. We would like to thank the editors of the volume, Bernard Comrie and Lucía Golluscio, for their insightful comments on the paper. We gratefully acknowledge the earlier contributions of Mily Crevels and Hein van der Voort through the Vidi project "Language diversity in the Guaporé region." Part of this paper was presented earlier at the Department of Linguistics at Stockholm University and in the LinC colloquium in Nijmegen.

they very different in nature? How do they stand in terms of validation against some independently established criterion?

In this paper we will take data from four areas of grammatical structure: argument marking (coded by Birchall), subordination (coded by Van Gijn), the noun phrase (coded by Krasnoukhova), and TAME marking (tense/aspect/mood/evidentiality, coded by Müller). These data are compared for 22 languages, thirteen from the Guaporé-Mamoré region in a broad sense, and nine from outside of the region. The key question we were originally asking ourselves is: do the thirteen languages from the region pattern more closely together than the overall set of languages as a whole, including the nine outsiders? It turned out that a somewhat different formulation was better, but we return to this below.

In section 2 the notion of linguistic area is further discussed, on the basis of the definition given in Thomason (2001), as applied to the Guaporé-Mamoré region, as well as some further methodological issues, and section 3 presents an outline of the current study. In section 4 the main quantitative results are given, using new measures, and section 5 concludes this paper.

It should be noted that there may not be so much difference between top-down and bottom-up approaches, as described above. Linguists may often work top-down to find areas, but they just do not do the bookkeeping. They validate the area with respect to controls, but in a very informal manner only. We propose a way of objectivizing this procedure and providing a quantitative basis for it, allowing significance testing.

2 The notion of linguistic area or Sprachbund applied to the Guaporé-Mamoré region

Campbell (2006) provides a comprehensive survey of the various definitions of 'linguistic area' or, alternatively Sprachbund (Trubetzkoy 1928). In the present paper, we take as a starting point Thomason's definition (2001: 99):

> ... a geographical region containing a group of three or more languages that share some structural features as a result of contact rather than as a result of accident or inheritance from a common ancestor.

This definition contains a number of key elements which call for independent justification for our perspective, when we want to apply it to the Guaporé-Mamoré region, its languages, and typical features, on the basis of the Crevels and Van der Voort (2008) study.

GEOGRAPHICAL REGION. The geographical region involved is the basin of the Guaporé and Mamoré rivers. As argued by Crevels and Van der Voort (2008), the region also shows shared cultural features, e.g., subsistence patterns, territorial subgroups with animal names. There is also considerable intermarriage between the different groups. There is also considerable shared cultural lexicon in words for e.g. 'maize', 'banana', 'bean', and 'chicha'.

THREE OR MORE LANGUAGES. The region is home to various languages from no less than 8 families and 11 isolates, thus totalling 19 different lineages.

SHARED STRUCTURAL FEATURES. A number of features are shared in the region such as a high incidence of prefixes, evidentials, directionals, verbal number, lack of nominal number, classifiers, and an inclusive/exclusive distinction. These features are fairly general, however.

CONTACT. Of course it remains to be established that the common features are due to contact, but there is considerable evidence for intensive contact between the languages in the region.

NOT ACCIDENT. As per tests carried out in the remainder of this paper.

INHERITANCE FROM A COMMON ANCESTOR. It is highly unlikely, given the present state of knowledge, that any or all of the languages in the area under consideration have a common ancestor demonstrable by similarities in basic vocabulary (Hammarström 2014 and references therein).

One of the issues to be considered here is how exactly to define and motivate the area under consideration. A strict geographical definition of the Guaporé-Mamoré region would exclude Cavineña, Moseten and Leko since their territories are part of the Beni river system and not the Mamoré. Karo would also be excluded. In order to include these languages, we would need to label the area as something different or at least be clear about the mismatch between the geographical terminology used for the region and our sample. The term Upper Madeira was used in Ramirez (2006) and Nimuendaju (1925) in reference to the geographical region, and not specifically a linguistic area. This term usually includes the Tupian culture area, which is technically not part of the Guaporé or Mamoré river systems, but rather of the Ji-Paraná/Machado system.

A similar issue concerns the exact locations of the Wichi and Tapiete varieties documented. The northern varieties are spoken near the upper reaches of the Mamoré headwaters and may have been in closer contact with Yurakaré and Moseten than, for example, Lakondê was in contact with Itonama. Considering these as completely distinct from the Guaporé-Mamoré sample may raise some bias in the analysis.

Eriksen (2011), using a large GIS-integrated data base, argues that there was ethnic specialization for trade in specific goods, and hence an incentive

to differentiate from neighbours. This may lead to conscious manipulation of language. Such a scenario predicts structural convergence but lexical divergence, since conscious manipulation typically targets lexicon but cannot access grammar.

It will be interesting to see if our data support the region as composed of multiple smaller areas (Lévi-Strauss 1948 defines three culture areas for this region: Moxo-Chiquitos, Chapacura, and Tupi) or if there is evidence for treating the region as an area without distinct subdivisions (like the Marico cultural complex argued for by Maldi 1991) – see Crevels and Van der Voort (2008) for discussion.

Three further methodological issues will need to be discussed before turning to our own research: the relation to genealogical historical linguistics, the predictive power of linguistic areas, and the issue of gradience.

The traditional perspective on linguistic areas is rooted in historical linguistics and the assumption of language families and crucially involves the notion of shared innovations, as can be seen in Table 1, based on Lindstedt (2000).

Tab. 1: The main grammatical features of the most important Balkan languages arranged implicationally. Table rearranged from the data in Lindstedt (2000).

	Balkan Slavic	**Albanian**	**Greek**	**Balkan Romance**	**Balkan Romani**
Object reduplication (= clitic doubling)	+	+	+	+	+
Relativum generale (= invariant relative clause marker)	+	+	+	+	+
Goal / location merger	+	+	+	+	(+)
Analytic comparison	+	+	(+)	+	+
Volo (= *want*) future	+	(+)	+	+	+
Dative / possessive merger	+	+	+	+	-
Past future as conditional	+	+	+	(+)	(+)
Enclitic articles	+	+	(+)	+	(+)
Evidentials	+	+	-	(+)	(+)
Habeo (= have) perfect	(+)	+	+	(+)	-
AUX(+COMP) + finite verb	+	(+)	+	(+)	+
Prepositions instead of cases	+	(+)	(+)	(+)	(+)

In the Balkans, the languages that participate in the linguistic area all belong to families (ultimately all part of Indo-European) also spoken outside the region: Balkan Slavic, Balkan Romance, Balkan Romani, or languages for which older material is available: Albanian, Greek. The branches or varieties of the present day Balkan region have all undergone changes that are not found in the earlier varieties or varieties outside of the region. Crucial then is the fact that there have been shared innovations.

The same holds for the postulated linguistic area of the Vaupés, where speakers of Arawakan (Aikhenvald 2002), Tucanoan, and Nadahup (Epps 2007) languages have interacted for a long time. All three groups have maintained their languages as separate entities, at least as far as the lexical shapes are concerned, as well as almost all morphemes; there has been little lexical borrowing. The Arawak language directly influenced by Tucano is Tariana, but other Arawak languages are not. Of the Nadahup languages, Hup has been affected by Tucano, but a language slightly further away, Dâw, much less, and Nadëb not at all. Thus the linguistic area is not simply postulated because features are shared, but because there are other languages of the same families that do not share the areal features.

In the Guaporé-Mamoré, in contrast, there are mostly isolates and small families, and other ways need to be found to argue that there is reason to assume that the languages in the region form a linguistic area or Sprachbund.

This brings us to the second methodological issue. Masica (2001) and Matras (2011) warn that assembling lists of shared features for areally adjacent languages has no explanatory value unless chance can be ruled out in favour of a specific contact scenario. That is, any set of languages may share a certain amount of features, to some degree, just by chance alone, and such cases obviously do not warrant an explanation in terms of language contact. In the present study, we explicitly test the amount of shared features for statistical significance.

As regards gradience, we may return again to Lindstedt (2000) for an illustrative example in his "Balkanization factor"; each Balkan language receives a score proportional to the number of features shared in the Balkan Sprachbund. In Table 1, the languages are arranged in terms of their degree of Balkanization, with Balkan Slavic being most Balkanized and Balkan Romani least Balkanized. Thus a Sprachbund may exhibit gradience in language membership, from "core" languages in the Sprachbund to "non-core" languages. This situation probably is found elsewhere as well, although it is rarely discussed in such explicit terms. Thus the Balkans can be characterized as a Sprachbund with "soft" boundaries: there is no sharp transition from Sprachbund to non-Sprachbund.

There is also gradience in features: "core" features in the Balkans would be Object reduplication or clitic doubling and a *Relativum generale* or the use of an invariant relative clause marker. In contrast, "non-core" features would be AUX(+COMP) + finite verb constructions and the use of prepositions instead of morphological case endings. Again, gradience in features is something found in many linguistic areas.

It is a challenge to build either kind of gradience into a quantitative model. In this paper, we perform tests that allow for gradience in the Sprachbund boundary and once a Sprachbund is identified, all features can be graded as to how well they characterize the Sprachbund in question.

Quantification of linguistic areas is a new field of study, though work in dialectology (Grieve et al. 2011 and references therein) has been targeting related questions about areal distributions of linguistic features. Grieve et al. (2011) discuss the problem of dividing a given region into dialect areas given a set of features and their geographical distribution. The problem of finding a linguistic area also starts out from a set of features and their geographical distribution, but is different from that of dialect area division (as defined in Grieve et al. 2011) for a number of reasons. For example, a dialect area requires a conglomerate of features whereas a linguistic area requires a conglomerate of features *which are due to contact*. Furthermore, a given area that needs to be dialect divided ultimately stems from one single (proto-)language meaning that all feature values of the dialects derive from the specific values the proto-language had. This is valuable information that can be used as a model constraint. In the case of a given area that needs to be searched for linguistic areas, there is no such guarantee, i.e., we cannot assume that all, or even a subset of, the features derive from a specific value of a proto-area or the like.

The first (and, to our knowledge, *only*) paper to directly address the question of finding linguistic area(s) given a set of features and their geographical distribution is Daumé (2009), who develops a Bayesian framework for linguistic-area searching and applies it to data in the World Atlas of Linguistic Structures (Haspelmath et al. 2005). In short, it describes a model for generating the observed data as a constrained mix of inherited, contact-induced and random features. It then searches heuristically for appropriate parameters for this mix that are highly likely given the actual WALS data. Among these parameters are the number of and locations of linguistic areas. This model is actually unnecessarily complicated for the dataset used in this paper. In our case, we can disregard genealogical inheritance (since the bulk of the languages of the Guaporé-Mamoré area are not related genealogically) and we can afford to do exhaustive searching. Furthermore, it has a few design choices, such as the use of the Pitman-Yor process to constrain the number of linguistic areas, that appear to have no other motivation than mathematical elegance.

None of the previous approaches allow for areas with "soft" or "gradient" boundaries, which is the focus of one of the test batches developed in the present paper.

3 Design of the current study

In this section we will outline the various features coded in our study. In section 3.1 argument marking in described, in 3.2 subordination, in 3.3 the noun phrase, and in 3.4 TAME marking.

3.1 Argument marking (coded by Birchall)

The questionnaire on argument marking codes for 80 structural features related to the way that core, oblique and derived arguments are encoded in the declarative main clause morphosyntax of the languages in our sample. Following standard practice in modern comparative linguistics, we start by examining the morphosyntactic treatment of prototypical semantic roles across different construction types, focusing on major distinctions such as intransitive subject (S), transitive agent (A), transitive patient (O) and ditransitive recipient (R). In some languages, further argument types are also considered, such as in split intransitive systems.

The primary areas of argument marking explored in the questionnaire include:

- Constituent order for S, A, O, R and obliques
- Head-marking of core arguments on the predicate, along with the grammatical features realized through markings such as person, number, gender and clusivity.
- Dependent-marking of core arguments and whether animacy plays a role in the realization of case markers.
- Treatment of oblique arguments and the number of formal distinctions made between them.
- The alignment of different argument types across various coding properties, and whether alignment patterns vary across different syntactic types of arguments (nominal vs. pronominal) or different tense and aspectual values.
- The presence of split intransitive and inverse alignment systems and their major typological parameters.

– Valency changing operations, with focus placed on their marking in the clause, the treatment of their derived arguments, and the major typological and semantic distinctions within the construction types.

Each feature is based on independent structural criteria, and the feature values coded for each language can be empirically verified by data available in our consulted sources and do not rely solely on the labels used by the various authors.

3.2 Subordination strategies (coded by Van Gijn)

The questionnaire on subordination strategies as it is used in this study[2] connects a number of semantically defined fields to the presence or absence of individual morphosyntactic features in the construction(s) encoding these semantic fields.

The choice of semantic fields or independent variables is basically determined by three factors: i) they are an adapted subset of the categories used by Cristofaro (2003), which allows for a comparison of the South American patterns with the global patterns found by Cristofaro; ii) they should yield a reasonably good representation of the subordination strategies a language uses, meaning that semantic relation types are chosen that are expected to yield different results, iii) information should be available from grammars for the majority of them, restricting the categories to the most common types. These considerations have led to the semantic types as given in Table 2.[3]

If a language in the sample has a construction that can be used to encode one or more of these semantic fields, it is taken into consideration. By taking this approach, we go well beyond the classic or canonical conception of subor-

2 The questionnaire on subordination strategies used for this paper is an adapted version of a much larger questionnaire developed by Rik van Gijn. In the original set-up of the questionnaire, the units of comparison are constructions rather than languages. In order to make the results for subordination more comparable to the other linguistic features discussed in this paper, an adaptation was made by which the systems of subordination strategies of languages as a whole are compared, and in which the number of data points per language is significantly reduced and less refined, for instance in the number of semantic relation types considered. The original approach as well as the algorithm that derives language signals from the constructions are explained and defended more thoroughly in Van Gijn and Hammarström (in preparation).

3 SoA stands for State of Affairs, defined as entities that "can be located in relative time and can be evaluated in terms of their reality status" (Hengeveld and MacKenzie 2008: 166). The subscripts 'M' and 'S' stand for 'main' and 'subordinate', respectively.

Tab. 2: Semantic relation types for subordination strategies.

Temporal (simultaneous, anterior)	SoA_S places SoA_Min a temporal perspective, indicating that SoA_M takes place at a moment after or overlapping with SoA_S.
Purpose	SoA_M is carried out in order to bring aboutSoA_S.
Phasal (terminative)	SoA_M indicates that some entity discontinues the temporal development of SoA_S in which s/he is involved as an agent.
Perception (see)	SoA_M expresses an act of visual perception; SoA_S expresses the state or action which is perceived.
Relatives (S, A, O)	SoA_S restricts the reference of some entity that is involved in SoA_M by describing a situation in which this same entity is involved as an A, S, or O argument.

dination, including clause combinations, nominalizations, non-finite clauses, serial verb constructions, auxiliary verb constructions, verb-verb compounds and derivational affixes (e.g. desiderative or causative markers). For this reason, we use the term subordination strategies rather than subordination.

The questions in the questionnaire concern the morphosyntactic encoding of the different semantic relation types. They have the general format "For semantic relation type X, is there a construction for which Z is true?", Z referring to a particular morphosyntactic characteristic. The answer possibilities are yes, no, do not know, and does not apply. Based on previous typological work (especially Lehmann 1988, Cristofaro 2003, Malchukov 2006 and Bickel 2010), four major concerns guide the questions in the questionnaire:

- Finiteness/deverbalization: it is often the case that subordinate or dependent predicates have fewer inflectional possibilities than independent or superordinate predicates. The questions that relate to this parameter are meant to determine what verbal categories can be marked independently on the subordinate event-denoting unit (EDU). Chosen variables for this version of the questionnaire are subject agreement and tense.
- Nominalization: questions relating to this parameter determine how noun-like the EDU is. Chosen variables: the (im)possibility to take case markers/adpositions, and the possibility of marking the subject as a possessor. Nominalization as a subordination strategy is one of the areal features of the Guaporé-Mamoré area mentioned by Crevels and Van der Voort (2008).
- Flagging: subordinate EDUs may or may not have a dependency marker associated with them (a marker that is added to the verbal or nominal inflection of the EDU mentioned above, and that does not occur on independent verbal predications). Questions that relate to dependency marking

concern the morphological status of the marker (bound or free). A special question concerns whether languages in the sample can mark switch reference on their temporal clauses, as this is also one of the Guaporé-Mamoré features noted by Crevels and Van der Voort (2008).
- Linearization: this pertains, for the version used in this study, only to the position of relativized NPs with respect to their restricting EDU.
- Integration: some of the semantic fields may be encoded by tighter constructions such as serial verb constructions, auxiliary verb constructions, verb-verb compounds and even derivational affixes. This is one of the areas in which a construction-based approach can yield more precise results. In the adaptation for this paper, three levels of integration are discerned: combinations of independent EDUs, constructions where the subordinate and superordinate EDUs are separate, but obligatorily contiguous words, and morphologically bound combinations of EDUs (affixes and V-V compounds).

The connection of the semantic relation types with the morphosyntactic features yields 38 data points per language.

3.3 The noun phrase (coded by Krasnoukhova)

The questionnaire is constructed in order to profile the structure of the noun phrase (NP) in the languages in the sample. We use the following working definition of an NP: a series of words, with a noun as its central constituent, which behaves as a single syntactic unit, and typically functions as an argument in a clause.

The NP questionnaire consists of 47 main questions and 29 sub-questions and covers such aspects as:
- Constituent order within the NP. Four modifier categories are taken into account: demonstratives, lexical possessors, numerals, and adjectives. All these are approached as semantic categories.
- Presence and realization of agreement within the NP. The agreement features considered are number, gender and physical properties, taking demonstratives, numerals and adjectives as potential agreement targets.
- Presence and conditions on the realization of number within the NP.
- Noun categorization devices, such as classifiers, and gender and noun class systems.
- Attributive possessive constructions. The parameters under investigation include: head vs. dependent marking of possession, the presence and for-

mal realization of the alienable/inalienable distinction, and the possessive strategies used by the languages in the sample.
– Spatial deixis, with a focus on semantic features that can be encoded by adnominal demonstratives.
– The availability of marking of temporal distinctions within the NP.

For some questions the answer values consist of 'yes', 'no' or 'not applicable', whereas for others, specific values are given as options. This concerns questions that cannot be answered adequately with an affirmative or negative response and require a more elaborate range of values. A detailed discussion and analysis of the NP features can be found in Krasnoukhova (2012).

3.4 TAME marking (Tense/Aspect/Mood/Evidentiality, coded by Müller)

The TAME questionnaire has four topic sections, one for each semantic category coded. All questions apply to non-negative, non-interrogative main clauses. Exceptions are the questions for negative imperative, and for purposive and irrealis which often occur in subordinate clauses. There are 38 questions distributed as follows:
– Tense has five (three independent, two dependent).
– Mood has 14 (all independent).
– Aspect has eight (all independent).
– Evidentiality has eight (seven independent, one dependent).

A dependent question can only be posed if one of the options in an independent question is realized. All independent questions can be answered in any order.

The questionnaire focuses on morphosyntactic marking including affixes, clitics, particles and auxiliaries (and for some questions, also repetition/reduplication). The questions differ in the way they have to be answered; the possibilities range from positive answers (yes, affix, clitic, particle, auxiliary, repetition, imperative marker plus negation, and values from zero to four in the case of remoteness distinctions) to negative (no) and neutral (unknown, not applicable). Each question has a value key that gives the possible answers.

The main challenge with TAME lies in the close connections between the categories and the need arises to specify each feature in a way that clearly distinguishes it from others. Markers can inherently belong to more than one category or acquire meanings similar to the original one, e.g. a future marker

might also be used for intentional and/or irrealis and the other way around. A present marker might be used for progressive and general truths, and a direct/visual evidential marker can have certainty values. In order to untangle the various meanings the questionnaire applies the "dominant parameter" and "prototype" principle introduced for typology by Dahl (1985), which means that a marker is entered under the question applying to its predominant meaning. If a marker has more than one major meaning of equal weight it is entered under every feature that applies to it.

The questions are designed to cover all major TAME features but do not go into fine detail. The tense section asks specifically for present, past and future tense as well as for remoteness distinctions for past and future. The mood/modality section asks for realis/irrealis, imperative (split into several questions according to person and number), prohibitive, intentional, potential, certainty/uncertainty (dubitative), frustrative, purposive, and desiderative. The aspect section asks for perfective/imperfective, perfect, habitual, continuative, iterative, and completive/incompletive. The evidentiality section asks for firsthand versus secondhand information (reportative) and quotative, third hand information, visual, inference and assumption.

All sections can easily be expanded, but the current goal is to present a broad picture of which categories are featured in a language and therefore categories with the best comparative values were chosen. For a more detailed discussion on coding TAME features see Müller (2013).

Appendix 1 illustrates the coding scheme adopted in this study.

3.5 Motivation for the language sample

A group of 22 mostly unrelated languages are studied here, of which 13 belong to the Guaporé-Mamoré river in a wide sense, and nine further away (although some are spoken in the Chaco region adjacent to the area under consideration). We have chosen these languages both to maximize genealogical diversity and because good grammars are available for them. Appendix 2 lists the languages in the present sample.

A total of 161 logically independent features are coded for these languages, as described in the previous sections.

4 A quantitative approach to linguistic areas

In this section we will explore some possible quantitative definitions of linguistic area. There are at least two points of contention that are left open within the definition of Thomason, namely:

- *soft/hard boundary*: Does the area have a hard boundary around it or does it decay in a more continuous manner?
- *convergence/cherry-picked*: Is the area such that the entire typological profiles have converged or is the number of shared features small relative to the total number of features in the languages? For example, if the Balkan Sprachbund is of the convergence type it makes sense to speak of a 'Balkan type' language, whereas if the Balkan Sprachbund only entails sharing a small number of features, members could be typologically very diverse on the whole, apart from those features.

We will first test hypotheses of convergence (sub-)area(s) in the Guaporé-Mamoré with a soft boundary. Any extant hard boundary Sprachbund would be expected to show up with a very non-linear, step-like, decay curve in our tests for soft boundary Sprachbunds. We informally inspect the curves accordingly rather than performing a separate explicit test.

Second we will test hypotheses of cherry-picked Sprachbund(s) with hard boundaries. To cherry-pick a set of features from a feature pool is computationally tractable if the area boundaries are hard, but we are unaware of a computationally tractable way to cherry-pick in our formalization of soft boundaries.[4]

In both cases, we will define a Bund-factor as a score for how well a given area(+features) tends to conform to the Sprachbund definition. We then enumerate all geographically coherent subsets (= areas) of the 22 languages and score them. For soft boundary convergence-Sprachbunds, enumerating all areas amounts to enumerating all centrepoints, i.e., the 22 languages. For hard-boundary cherry-picked Sprachbunds, enumerating all areas amounts to enumerating all centrepoints multiplied by all sizes, i.e., 22 languages * 21 sizes = 462 areas.

Since nearly all the languages in question are genealogically unrelated, genealogical inheritance can hardly be responsible for any major patterns in the data. We use permutation tests to check the areality of patterns of the

4 When picking with a hard boundary, it is always optimal to pick the feature with the highest agreement for the area considered, regardless of which features have so far been picked. In the case of soft boundaries, there may be dependencies, which makes the picking process highly non-trivial.

shared features, ruling out chance as an explanation. Using permutation tests to check for areality obviates the need to consider potential universal functional dependencies[5] among features – if there are such dependencies they should be reflected uniformly in areally coherent versus non-areally coherent subsets of languages. Thanks to a number of control languages outside the Guaporé-Mamoré area, the present test set is able to distinguish whether the Guaporé-Mamoré area as a whole is a big Sprachbund or whether the languages there reflect universal principles in language design. Any remaining areally patterned sharing of features must thus be due to contact, as required in the Sprachbund definition.

4.1 Soft-Boundary Convergence Sprachbunds

Assuming a centre point language *l* and an area with a soft boundary, an area is defined by that centre point *l* and a decay parameter *k*, the idea being that the typological distance from *l* should increase with geographical distance. If there is a sharp increase in typological distance at some geographic distance, then a hard boundary can be said to exist, otherwise not.

First, it is necessary to give an operational definition of Typological Distance. We will use the simplest kind of measure that assumes that all features are of equal weight, namely the Hamming distance:

$$\text{Given two languages, their typological distance is} = \frac{\text{the number of features where the languages have a different value}}{\text{the number of features where both languages have a value}}$$

For example, between Cavineña [cav] and Mekens [skf], there are 125 (out of the total 164) features which are defined for both languages. Of these, they differ in 56 features so their distance is $56/125 \approx 0.448$.

Note also that no distinction is made between feature agreement as to the absence of a feature and agreement as to presence of a feature. For example, one question might be 'has nominal number' – if both languages lack it, i.e. has a N-value, it counts as much as when both languages have it, i.e. has a Y-value. This is consistent with the methodology of Crevels and Van der Voort (2008) who count, for example, lack of nominal number as a shared feature.

Now, for example, consider Cavineña (see Table 3). If there is a Sprachbund centered at Cavineña, we would expect the typological distance to increase as we get further away from Cavineña (see Figure 1 for a plot).

5 Areal-specific functional dependencies (cf. Dunn et al. 2011) are probably not distinguishable from Sprachbunds.

Tab. 3: Example geographical and typological distances from Cavineña.

	Movima	Mosetén	Leco	Itonama	Kanoé	Baure	Wari'
Geographical distance (km)	118.7	193.9	228.3	253.4	257.6	265.7	371.0
Typological distance	0.508	0.512	0.400	0.555	0.457	0.621	0.632

We may try to fit a line using standard linear regression to show how much of a relationship there is, giving a slope k and the Pearson correlation coefficient r measuring how well the line fits. In the case of Cavineña, typological distance does not increase with geographical distance. Also, there is nothing to suggest a non-linear relationship, as in Figure 1.

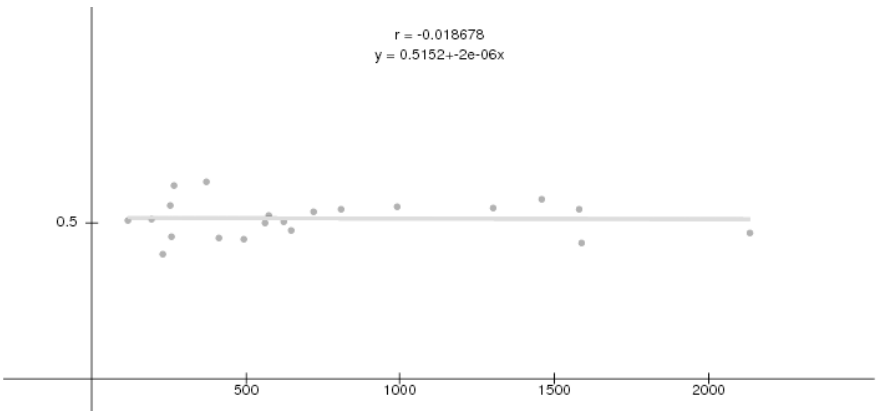

Fig. 1: A plot of geographic distance (x-axis) versus typological distance (y-axis) for the potential Sprachbund centred at Cavineña.

Table 4 below shows the slope k and fit r for each centrepoint language. Intuitively, none except the top three show even a weak tendency. To test which tendencies are stronger than random, we generate 1000 random worlds by permuting the locations of the languages. For a given centrepoint and k,r in the real world, we check in how many of the random worlds the corresponding centrepoint exhibits a higher r value. Only the Wichí Lhamtés Nocten- and Pilagá-centered Sprachbunds, outside the Guaporé-Mamoré, approach statistical significance ($p \approx 0.002$ and $p \approx 0.012$ respectively). But these p-values might not stand-up under correction for multiple (22 cases) testing. The result

Tab. 4: Potential Sprachbunds and their tendencies to exhibit a linear geographical-typological relationship. Pearson's *r* measures the goodness of fit of the linear relationship and *k* the actual slope of the best fitting line.

Sprachbund Centre	*K*	*r*
WichíLhamtésVejoz [wlv]	6.82e-05	0.564
Pilagá [plg]	5.43-05	0.498
Desano [des]	3.98-05	0.452
Tapiete [tpj]	3.23e-05	0.293
Hupdë [jup]	2.69e-05	0.253
Cusco Quechua [quz]	3.35e-05	0.176
Central Aymara [ayr]	2.19e-05	0.166
Sakirabiá [skf]	3.49e-06	0.032
Kanoé [kxo]	1.96e-06	0.021
Leco [lec]	1.51e-06	0.011
Mosetén-Chimané [cas]	−6.18e-07	−0.006
Trió [tri]	−9.40e-07	−0.010
Cavineña [cav]	−1.97e-06	−0.019
Itonama [ito]	−2.57e-06	−0.025
Movima [mzp]	−3.09e-06	−0.038
Baure [brg]	−7.14e-06	−0.052
Lakondê [lkd]	−1.01e-05	−0.077
Yuracaré [yuz]	−7.41e-06	−0.133
Apurinã [apu]	−2.35e-05	−0.184
Wari' [pav]	−1.87e-05	−0.185
Kwazá [xwa]	−2.65e-05	−0.201
Karo [arr]	−2.27e-05	−0.203

for the Guaporé-Mamoré area is clear in the present dataset: there is no evidence for a soft-boundary convergence Sprachbund in the area. Also, although omitted for reasons of space, none of the decay curves suggest the existence of a hard-bounded convergence Sprachbund.

4.2 Hard-Boundary Cherry-Picked Sprachbunds

A hard-boundary area *A* can be specified in terms of centre language *l* and a size *n* as 'the *n* geographically closest neighbours of *l* (including *l*)'. For example, the three nearest neighbours of Kwazá are Karo, Mekens, and Lakondê, so A(Kwazá, 4) = {Kwazá, Karo, Mekens, Lakondê}. This means than on *n* languages there are no more than *n*(*n*-1) hard-boundary areas, or 22 * 21 = 462 in the present study.

Given an area $A(l,n)$ and a (sub)set of features F we can calculate its agreement as:

$$AGR(A(l,n),F) = \frac{1}{|F|}\sum_{f\in F} max\_fraction(A(l,n),f)$$

Where max_fraction gives the fraction of the most popular feature value for a given feature in a set of languages.[6] For example, for a given feature *f*, if Kwaza and Karo have the value X, Mekens has Y and Lakondê has Z, then the most popular value is X, and its fraction is 2/4 = 0.5.

The first important observation is that the mere *number* of features (as opposed to *what* the features are) is sufficient to calculate the maximal agreement for a given area. If we have to select *n* features from a feature pool yielding maximal agreement, we simply have to rank the features as to max_fraction and select the *n* top ones.

Figure 2 exemplifies how agreement depends on number of features for Kwazá and its two nearest neighbours. For such a small set of languages, it is easy to get 100 % agreement when selecting a small number of features from a large pool of features. However, the more features that need to be selected, the more accuracy has to be sacrificed. With a larger set of languages, high accuracy is naturally more difficult to achieve at the corresponding levels.

As Figures 3 and 4 below show, different potential areas of the same size have different agreement trajectories, and different areas of the same size can be straightforwardly compared. But to compare the agreement trajectories for areas of different sizes we need to compare how *non-random* they are.

This leads us to the second important observation: given an area there is an optimal number of features in the sense that the fewest random worlds exhibit a higher level of agreement. For a given area and its agreement trajectory in the real world, we can count in how many of 1000 random worlds the real world agreement beats the agreement in the corresponding random-world area and number of features. As exemplified in Figure 5, agreement starts out high and decreases subsequently, while non-randomness starts out low (the agreement level is matched by almost all random worlds) and reaches a maximum at some higher number of features. Thus for each area we can calculate a specific optimal number of features by taking the value that beats the most random worlds.

6 Note that agreement (AGR) is not the same as inverted Hamming distance. Agreement can never be less than 0.5 for a binary feature (in general, never less than 1/n for a feature with n possible values). For example, if 11 has feature vector [X Y] and 12 has [X Z] their Hamming distance is 1/2 but their agreement is (1.0 + 0.5)/2 = 0.75.

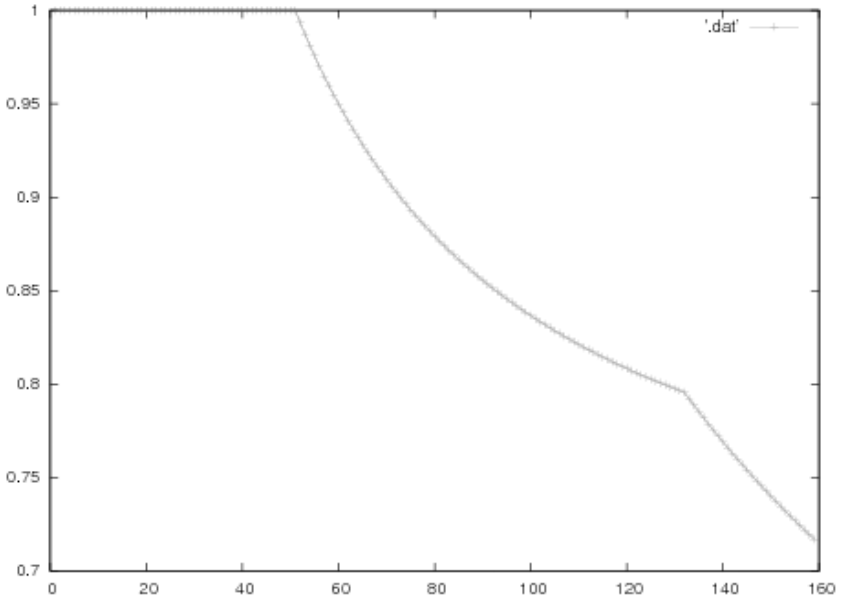

Fig. 2: Agreement for the set of languages {Kwazá, Karo, Mekens} as it depends on the number of features to be picked.

Now we are ready to test the 462 possible areas.[7] Again 1000 random worlds are generated and we check how many of those random worlds beat the real world areas in terms of their $A(l,n)$ and optimal # features. Some 41 out of the 462 areas turn out to be significant at the conventional $p < 0.01$ level. They are shown in Appendix 3, grouped by centrepoint. The p-values should not be taken at face value since they have not been corrected for multiple testing (this would require a deeper investigation of random worlds than was possible for the present study).

The top three centres shown in the table in Appendix 3 are outside the Guaporé-Mamoré-centered areas. Interesting as they may be, they are not pursued here as our current dataset is not dense enough with languages from and around those areas.

Of the remaining areas in the Guaporé-Mamoré, some cells in the table in Appendix 3 actually denote the same area, i.e. set of languages. For example, Kanoé's 3 nearest neighbours are Itonama, Baure and Wari', and Waris 3 near-

7 Note that we do not test the significance of all (A(l, n), #f) pairs as that would be so many tests that it would require far more random worlds to check significance.

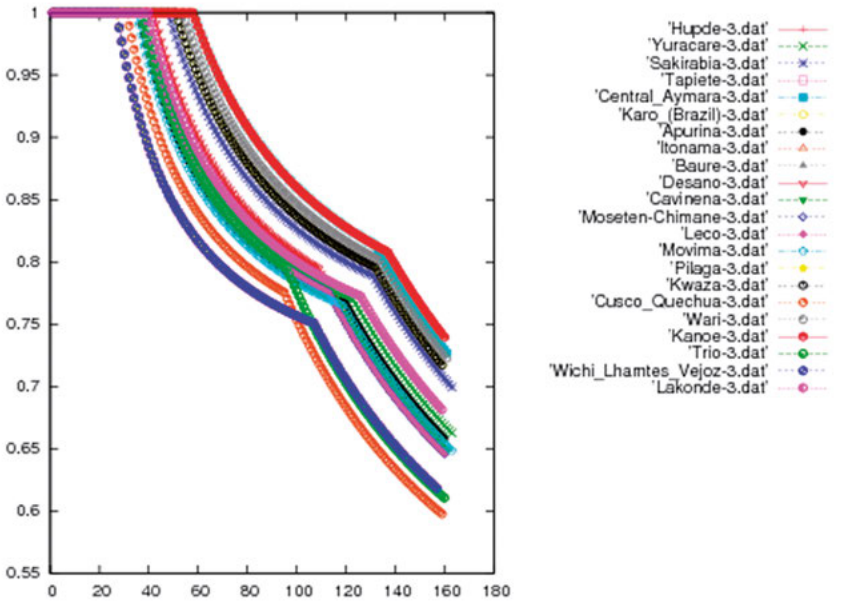

Fig. 3: Agreement trajectory for areas of size 3.

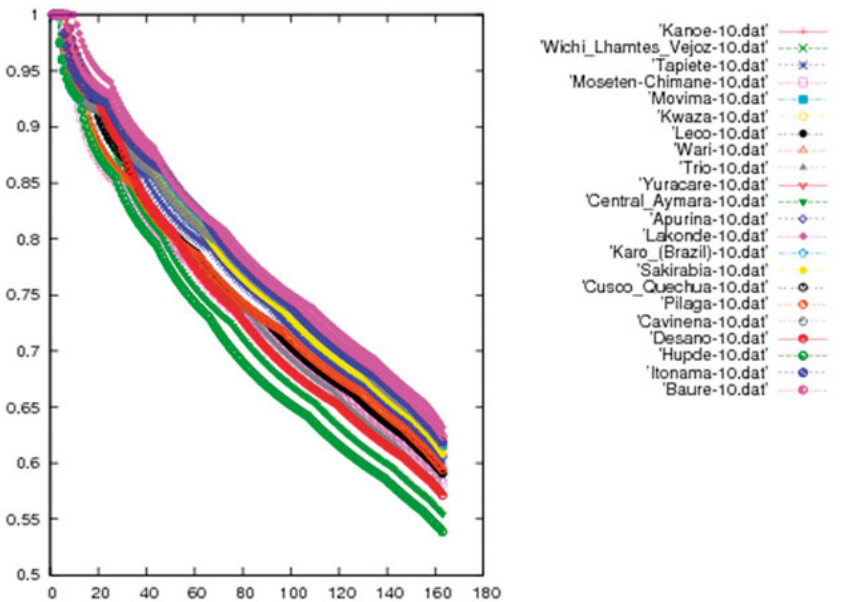

Fig. 4: Agreement trajectory for areas of size 10.

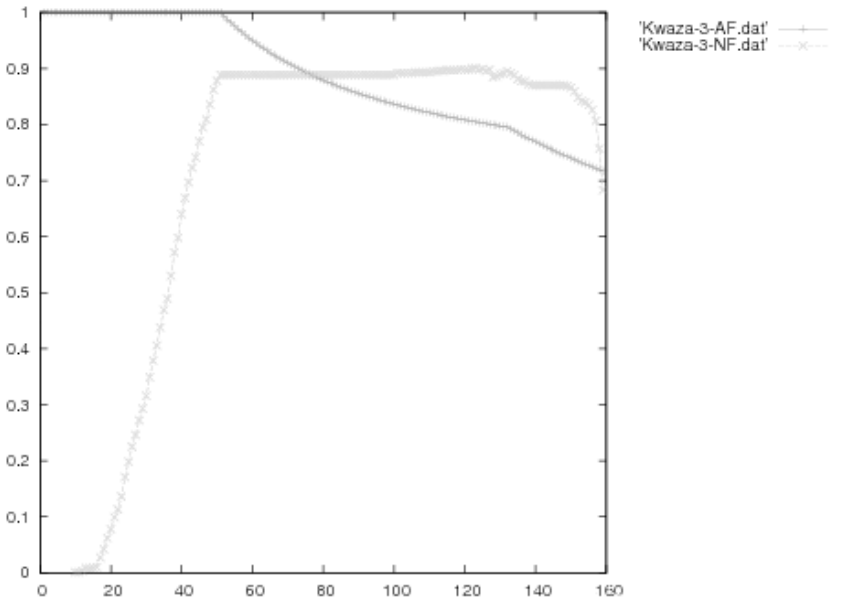

Fig. 5: Agreement and fraction of random worlds with lower agreement, as functions of # features to be cherry-picked, for the set of languages {Kwazá, Karo, Mekens}.

est neighbours are Kanoé, Itonama, Baure, so Kanoé-4 and Wari'-4 are the same area. Other cells are epiphenomenal in the sense that they are extensions of other cells which are more significant. For example, Kanoé-4 has a significance of .005, and Kanoé-5 – which is just Kanoé-4 with Movima added to it – has less significance (.009). Thus adding Movima made the whole less of a Sprachbund and it is only significant, albeit diluted, because the original Sprachbund was. Similarly, given a very significant Sprachbund, if we take away one language that makes it less significant, it may still be quite significant if the original Sprachbund was very significant. Such cases – diminished variants of more significant cells – are also epiphenomenal. If we remove all duplicate cells, epiphenomenal cells, and non-Guaporé-Mamoré cells, only two Sprachbunds remain, as show in Table 5.

Intuitively speaking, the remaining two Sprachbunds, without being fully epiphenomenal, are in fact also quite similar in that 4 languages are shared between the two. They can be taken as alternative answers to Crevels and Van der Voort's (2008: 166, 172) question about the appropriate boundaries and break-up of the Guapore-Mamoré linguistic area.

Which are the features that account for the Sprachbunds in Table 5? Appendix 4 shows the 55 cherry-picked features and agreement levels of the most

Tab. 5: The two hard-bounded Sprachbunds in the Guaporé-Mamoré area remaining after purging. The table cells show the *p*-value and number of cherry-picked features.

	45	**78**
Baure-5	**0.000**	
{Baure, Itonama, Kanoé, Movima, Mekens}	55	
Karo-8/Kwazá-8/Mekens-8/Lakondê-8		**0.003**
{Lakondê, Kwazá, Mekens, Karo, Baure, Itonama, Kanoé, Wari'}		100

significant Sprachbund Baure-5 (for comparison, the corresponding agreement levels for the Karo-8 Sprachbund are shown in parallel).

The 55 cherry-picked features come from all four subdomains, but with a dominance of TAME features (Table 6). Inspection shows that in fact 16 out of the 18 cherry-picked high-agreement TAME features are agreement in negative values. A large proportion of the agreement in the Baure-6 Sprachbund is thus accounted for by the shared *lack* of certain feature values.

Tab. 6: Domain and agreement distribution of the 55 cherry-picked features for the Baure-5 Sprachbund.

	Number of cherry-picked features	**Total number of features**	**Percentage**	**Agreement**
TAME	18	36	50.0 %	0.93
NP	17	54	31.5 %	0.94
ARG	14	56	25.0 %	0.90
SUB	6	18	33.3 %	0.93

The 55 cherry-picked features are not directly comparable to the 24 features considered by Crevels and Van der Voort (2008), which have a different granularity and include features from other domains such as pronominal system and phonology. However, it is still clear that most of the cherry-picked features have no counterpart in Crevels and Van der Voort (2008: 166–172).

5 Discussion, conclusions and suggestions for further research

In this paper we have tested several approaches for quantifying measures relating to Sprachbund phenomena, using data from the Guaporé-Mamoré area in central South America. A number of conclusions can be drawn.

There was little evidence for structural convergence of *entire typological profiles*. However, there is good evidence for the convergence of *some* features due to contact, and the cherry-picking Sprachbund measure allows us to pinpoint which features are shared and exactly which languages are involved in a statistically optimal language * feature mix.

Using this measure, we come closer to the results reached by Crevels and Van der Voort (2008).

There are, however, a number of possibly controversial choices to be made when quantifying linguistic and areal data. In the present approach we have assumed that all features and feature values for each feature have equal weight, that language locations can be approximated with a centrepoint, that distances between languages can be approximated by distance as the crow flies, that Sprachbunds should be round (rather than e.g. oval, or shaped by some territorial features) and so on. Results are subject to change drastically if all or any of these crude measures are improved.

Much will depend, also, on the precise feature set taken into consideration using any methodology.

References

Aikhenvald, Alexandra Y. 2002. *Language contact in Amazonia*. Oxford: Oxford University Press.

Bacelar, Laércio Nora. 2004. Gramática da língua Kanoê. Katholieke Universiteit Nijmegen Ph.D. dissertation.

Bickel, Balthasar. 2010. Capturing particulars and universals in clause linkage: A multivariate analysis. In Isabelle Bril (ed.), *Clause linking and clause hierarchy*, 51–101. Amsterdam: John Benjamins.

Campbell, Lyle. 2006. Areal linguistics: A closer scrutiny. In Yaron Matras, April McMahon & Nigel Vincent (eds.), *Linguistic areas: Convergence in historical and typological perspective*, 1–31. Basingstoke: Palgrave MacMillan.

Carlin, Eithne. 2004. *A grammar of Trio. A Cariban language of Suriname.* Frankfurt/Main: Peter Lang.

Cerrón-Palomino, Rodolfo & Juan Carvajal Carvajal. 2009. Aimara. In Mily Crevels & Pieter Muysken (eds.), *Lenguas de Bolivia. Tomo I: Ámbito andino*, 169–213. La Paz: Plural editores.

Crevels, Mily & Hein van der Voort. 2008. The Guaporé-Mamoré region as a linguistic area. In Pieter Muysken (ed.), *From linguistic areas to areal linguistics* (Studies in Language Companion Series 90), 151–179. Amsterdam/Philadelphia: John Benjamins.

Crevels, Mily. 2006. Verbal number in Itonama. In Grazyna J. Rowicka & Eithne B. Carlin (eds.), *What's in a verb?* (LOT Occasional Series 5), 159–170. LOT: Utrecht University.

Crevels, Mily. 2012. Itonama. In Mily Crevels & Pieter Muysken (eds.), *Lenguas de Bolivia. Vol. 2. Amazonía*, 233–296). La Paz: Plural editores.
Cristofaro, Sonia. 2003. *Subordination*. Oxford: Oxford University Press.
Dahl, Östen. 1985. *Tense and aspect systems*. Oxford: Basil Blackwell.
Danielsen, Swintha. 2007. *Baure. An Arawak language of Bolivia* (Indigenous Languages of Latin America 6). Leiden: CNWS Publications. Radboud Universiteit Nijmegen Ph.D. dissertation.
Daumé, Hal, III. 2009. Non-parametric Bayesian areal linguistics. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (NAACL '09), 593–601. Morristown, NJ, USA: Association for Computational Linguistics.
Dunn, Michael, Simon J. Greenhill, Stephen C. Levinson & Russell D. Gray. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473. 79–82.
Epps, Patience. 2007. The Vaupés melting pot: Tukanoan influence on Hup. In Alexandra Y. Aikhenvald & R. M. W. Dixon (eds.), *Grammars in contact. A cross-linguistic typology*, 267–289. Oxford: Oxford University Press.
Epps, Patience. 2008. *A grammar of Hup*. Berlin: Mouton de Gruyter.
Eriksen, Love. 2011. Nature and culture in prehistoric Amazonia. Using GIS to reconstruct ancient ethnogenetic processes from archaeology, linguistics, geography, and ethnohistory. Lund: Lund University doctoral dissertation.
Everett, Dan & Barbara Kern. 1997. *Wari'. The Pacaas Novos language of Western Brazil*. London: Routledge.
Facundes, Sidney da Silva. 2000. The language of the Apurinã people of Brazil. SUNY Buffalo Ph.D. dissertation.
Gabas Jr., Nilson. 1999. A grammar of Karo, Tupí (Brazil). University of California, Santa Barbara Ph.D. dissertation.
Galucio, Vilacy. 2001. The morphosyntax of Mekens (Tupi). University of Chicago Ph.D. dissertation.
González, Hebe Alicia. 2005. A grammar of Tapiete (Tupí-Guaraní). University of Pittsburgh Ph.D. dissertation.
Grieve, Jack, Dirk Speelman & Dirk Geeraerts. 2011. A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change* 23. 1–29.
Guillaume, Antoine. 2008. *A grammar of Cavineña*. Berlin: Mouton de Gruyter.
Hammarström, Harald. 2014. Lexical relationships of South American languages. In Loretta O'Connor & Pieter Muysken (ed.), *The native languages of South America. Origins, development, typology*, 56–69. Cambridge: Cambridge University Press.
Hardman, Martha, Juana Vásquez & Juan de Dios Yapita Moya. 1988. *Aymara: compendio de estructura fonológica y gramatical*. La Paz: Ediciones ILCA (Instituto de Lengua y Cultura Aymara); Gainesville: The Aymara Foundation.
Hardman, Martha. 2001. *Aymara*. Munich: Lincom Europa.
Haspelmath, Martin, Matthew S. Dryer, David Gil & Bernard Comrie (eds.). 2005. *World atlas of language structures*. Oxford: Oxford University Press.
Haude, Katharina. 2006. A grammar of Movima. Radboud Universiteit Nijmegen Ph.D. dissertation.
Hengeveld, Kees & J. Lachlan Mackenzie. 2008. *Functional Discourse Grammar*. Oxford: Oxford University Press.

Krasnoukhova, Olga. 2012. *The noun phrase in the languages of South America*. Utrecht: LOT.

Lehmann, Christian. 1988. Towards a typology of clause linkage. In John Haiman & Sandra A. Thompson (eds.), *Clause combining in grammar and discourse,* 181–226. Amsterdam/ Philadelphia: John Benjamins.

Lévi-Strauss, Claude. 1948. Tribes of the right bank of the Guapore river. In J. H. Steward (ed.), *Handbook of South American Indians* 3, 370–379. Washington DC: Smithsonian Institution.

Lindstedt, Jouko. 2000. Linguistic Balkanization: Contact-induced change by mutual reinforcement. In Dicky Gilbers, John Nerbonne & Jos Schaeken (eds.), *Languages in contact* (Studies in Slavic and General Linguistics 28), 231–246. Amsterdam & Atlanta, GA: Rodopi.

Malchukov, Andrej. 2006. Constraining nominalization: form/function competition. *Linguistics* 44(5). 973–1009.

Maldi, Denise. 1991. O complexo cultural do Marico: Sociedades indígenas dos Rios Branco, Colorado e Mequens, afluentes do Médio Guaporé. *Boletim do Museu Paraense Emílio Goeldi, Antropologia* 7(2). 209–269.

Masica, Colin. 2001. The definition and significance of linguistic areas: methods, pitfalls, and possibilities (with special reference to the validity of South Asia as a linguistic area). In R. Singh, P. Bhaskararao & K.V. Subbarao (eds.), *The yearbook of South Asian languages and linguistics 2001. Tokyo symposium on South Asian languages: Contact, convergence and typology*. New Delhi: Sage, 205–267.

Matras, Yaron. 2011. Explaining convergence and the formation of Linguistic Areas. In Osamu Hieda, Christa König & Hirosi Nakagawa (eds.), *Geographical typology and linguistic areas: With special reference to Africa*, 143–160. Amsterdam: John Benjamins.

Meira, Sérgio. 1999. A grammar of Tiriyó. Rice University Ph.D. dissertation.

Miller, Marion. 1999. *Desano grammar.* Dallas: Summer Institute of Linguistics and the University of Texas at Arlington.

Müller, Neele. 2013. *Tense, aspect, modality, and evidentiality marking in South American indigenous languages*. Utrecht: LOT.

Nimuendaju, Curt. 1925. As tribus do Alto Madeira. *Journal de la Société des Américanistes de Paris*, n.s., 17. 137–172.

Ramirez, Henri. 2006. As línguas indígenas do Alto Madeira: Estatuto atual e bibliografia básica. *Língua Viva* 1(1). 1–16.

Sakel, Jeanette. 2004. *A grammar of Mosetén.* Berlin: Mouton de Gruyter.

Telles, Stella. 2002. Fonologia e gramática Latundê/Lakondê. Vrije Universiteit Amsterdam Ph.D. dissertation.

Terraza, Jimena. 2009. Grammaire du wichí: Morphologie et morphosyntaxe. Université du Québec à Montréal Ph.D. dissertation.

Van de Kerke, Simon. 2000. Case marking in the Leko language. In Hein van der Voort & Simon van de Kerke (eds.), *Ensaios sobre lenguas indígenas de las tierras bajas de Sudamérica: Contribuciones al 49º Congreso Internacional de Americanistas en Quito 1997* (Lenguas Indígenas de América Latina (ILLA) 1), 25–37. Leiden: Research School of Asian, African and Amerindian Studies (CNWS), Universiteit Leiden.

Van de Kerke, Simon. 2002. Complex verb formation in Leko. In Mily Crevels, Simon van de Kerke, Sérgio Meira & Hein van der Voort (eds.), *Current studies on South American languages* (Lenguas Indígenas de América Latina (ILLA) 3), 241–254. Leiden: Research School of Asian, African and Amerindian Studies (CNWS), Universiteit Leiden.

Van de Kerke, Simon. 2006. Object cross-reference in Leko. In Grazyna J. Rowicka & Eithne B. Carlin (eds.), *What's in a verb?* (LOT Occasional Series 5), 171–188. LOT : Utrecht University.

Van de Kerke, Simon. 2009. El Leko. In Pieter Muysken & Mily Crevels (eds.), *Ámbito andino* (Lenguas de Bolivia 1), 287–332. La Paz: Plural Editores.

Van der Voort, Hein. 2004. *A grammar of Kwaza*. Berlin: Mouton de Gruyter.

Van Gijn, Rik & Harald Hammarström. In prep. A construction-based approach to measuring distances between languages: Subordination strategies in South America. Ms.

Van Gijn, Rik. 2006. A grammar of Yurakaré. Radboud Universiteit Nijmegen Ph.D. dissertation.

Vidal, Alejandra. 2001. Pilagá grammar (Guaykuruan family, Argentina). University of Oregon Ph.D. dissertation.

Weber, David. 1996. *A grammar of Huallaga (Huánuco) Quechua*. Berkeley, CA: University of California Press.

Appendix 1: Example of the coding scheme adopted in the study

Language	Applicative construction can change the type of object the verb takes without changing valency	Are demonstratives related to third person pronouns?	Are there any (grammaticalized) sex-markers? (realized on the noun itself)	In possessive constructions with pronominal possessor, is the 'possessed' noun usually marked? (state most frequent construction; if depends on alienability, state marking in alienable constructions)	Person and number distinctions are morphologically conflated
Cavinena [cav]	?	c	0	a	?
Hupde [jup]	0	a	1	a	?
Pilaga [plg]	1	e	1	b	0
Trio [tri]	?	e	0	b	0
Wichi Lhamtes Nocten [mtp]	1	?	?	?	0

Appendix 2: The language sample

Language	Family	Country	Author	NP	TAME	ARG	SUB
Within Guaporé-Mamoré Area							
Baure	Arawakan, Bolivia-Parana	BO	Danielsen 2006	*	*	*	*
Aymara	Aymaran	BO/PE/CH	Hardman 2001, Hardman et al. 1988, Cerrón Palomino & Carvajal Carvajal 2009	*	*	*	*
Wari'	Chapacuran	BR	Everett & Kern 1997	*	*	*	*
Mosetén	Isolate	BO	Sakel 2004	*	*	*	
Cavineña	Tacanan	BO	Guillaume 2008	*	*	*	*
Itonama	Isolate	BO	Crevels 2006, 2012, p.c.	*	*	*	*
Lakondê	Nambikwaran	BR	Telles 2002				
Leko	Isolate	BO	Van de Kerke 1998–2006, 2009, p.c.	*		*	
Movima	Isolate	BO	Haude 2006	*	*	*	*
Yurakaré	Isolate	BO	Van Gijn 2006	*	*	*	*
Kanoê	Isolate	BR	Bacelar 2004	*	*	*	*
Mekens	Tupian, Tupari	BR	Galucio 2001	*	*	*	*
Kwazá	Isolate	BR	Van der Voort 2004	*	*	*	*

Language	Family	Country	Author	NP	TAME	ARG	SUB
Outside Guaporé-Mamoré Area							
Huallaga Quechua	Quechuan	PE	Weber 1996	*	*		*
Pilagá	Guaycuruan	AR	Vidal 2001	*	*	*	*
Hup	Nadahup	BR	Epps 2008	*	*	*	*
Trio/Tiriyó	Cariban, Taranoan	SU/BR	Carlin 2004, Meira1999	*	*	*	*
Wichí of Rivadavia (Mataco)	Matacoan	AR	Terraza 2009	*		*	*
Apurinã	Arawakan, Purus	BR	Facundes 2000	*	*	*	*
Desano	Tucanoan	CO	Miller 1999, Wilson Silva p.c.	*	*	*	*
Karo	Tupian, Ramarama	BR	Gabas Jr. 1999	*	*	*	*
Tapiete	Tupian, Tupi-Guarani	BO/AR	González 2005	*	*	*	*

Appendix 3: The 41 hard-bounded Sprachbunds grouped by centre and size

The table cells show the p-value and number of cherry-picked features.

	Size													
	4	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**
C Aymara	**.002** 45	**.001** 59												
Pilagá														**.002** 147
Tapiete														**.002** 147
Kanoé	**.005** 100	**.009** 132					**.009** 62			**.005** 67				
Movima									**.009** 72					
Itonama		**.009** 132	**.000** 49				**.009** 62	**.001** 57	**.009** 72					
Mekens				**.004** 46	**.003** 100	**.007** 53						**.000** 70		
Yuracaré								**.002** 58		**.003** 116	**.000** 133			

	Size															
	4	5	6	7	8	9	10	11	12	13	14	15	16	17		
Wari'	**.005**	**.005**														
	100	76														
Kwazá				**.004**	**.003**	**.007**						**.000**				
				46	100	53						70				
Karo				**.008**	**.003**	**.007**										
				87	100	53										
Baure		**.000**					**.008**	**.001**	**.009**							
		55					68	57	72							
Lakondê				**.004**	**.003**	**.007**	**.003**					**.000**	**.000**			
				46	100	53	134					70	51			

Appendix 4: The 55 cherry-picked features and agreement levels of the most significant Sprachbund Baure-5 and the corresponding agreement levels for the Karo-8 Sprachbund in parallel

		{Baure, Itonama, Kanoé, Movima, Mekens}		{Baure, Itonama, Kanoé, Karo, Kwazá, Lakondê, Mekens, Wari'}	
ARG	S is case-marked	*c*	1.000	*c*	1.000
ARG	O is case-marked	*c*	1.000	*c*	0.875
ARG	A is case-marked	*c*	1.000	*c*	1.000
NP	What is the most frequent order of relative clause and noun?	*b*	1.000	*b*	0.750
NP	Is there a number distinction in 3rd person pronouns?	*b*	1.000	*b*	0.875
NP	What system of demonstratives is present in the language? (distance-oriented system vs. person-oriented system)	*a*	1.000	*a*	0.625
NP	What is the most frequent order of numeral and noun?	*a*	1.000	*a*	0.500
NP	What is the most frequent order of demonstrative and noun?	*a*	1.000	*a*	0.875
NP	Are there indefinite articles in use?	*a*	1.000	*a*	1.000
TAME	Can or must time be expressed by lexemes or periphrastic constructions?	*Y*	1.000	*Y*	1.000
SUB	Temporal clauses with switch reference	*N*	1.000	*N*	0.750
SUB	Relative clauses with possessor subject	*N*	1.000	*N*	0.875
SUB	Perception relations encoded by an syntactically integrated structure	*N*	1.000	*N*	0.875
SUB	Phasal relations encoded by an syntactically integrated structure	*N*	1.000	*N*	0.875
TAME	Is visual evidentiality marked morphologically?	*N*	1.000	*N*	0.750
TAME	Is there an inferred evidentiality morphological marker?	*N*	1.000	*N*	0.750
TAME	Is there an assumed evidentiality morphological marker?	*N*	1.000	*N*	0.875
TAME	Is purposive marked morphologically?	*N*	1.000	*N*	0.875

		{Baure, Itonama, Kanoé, Movima, Mekens}		{Baure, Itonama, Kanoé, Karo, Kwazá, Lakondê, Mekens, Wari'}	
TAME	Is present tense marked morphologically?	*N*	1.000	*N*	0.875
TAME	Is non-firsthand information marked morphologically?	*N*	1.000	*N*	1.000
TAME	Is imperfective marked morphologically?	*N*	1.000	*N*	0.875
TAME	Is imperative for 3rd person only marked morphologically?	*N*	1.000	*N*	1.000
TAME	Is firsthand information marked morphologically?	*N*	1.000	*N*	1.000
TAME	Is an incompletive action marked morphologically?	*N*	1.000	*N*	1.000
TAME	Is a completive action marked morphologically?	*N*	1.000	*N*	0.875
NP	Are there nouns denoting obligatorily possessed items?	*1*	1.000	*1*	0.750
ARG	Language shows contrasting alignment patterns for S depending on tense and/or aspectual values of the verb:	*0*	1.000	*0*	0.875
NP	Do numerals receive any special (class-changing) morphology in order to function as an attributive modifier within an NP?	*0*	1.000	*0*	0.625
NP	Do nouns have a morphologically marked paucal?	*0*	1.000	*0*	0.875
NP	Do nouns have a morphologically marked dual?	*0*	1.000	*0*	1.000
NP	Do discontinuous NPs occur?	*0*	1.000	*0*	0.875
ARG	Causation can be expressed through verb serializations	*0*	1.000	*0*	0.875
ARG	Causation can be expressed through quotative constructions	*0*	1.000	*0*	0.875
NP	Can adnominal demonstratives encode altitude?	*0*	1.000	*0*	1.000
ARG	Applicative constructions can be formed through verb serialization	*0*	1.000	*0*	0.625
TAME	Is there a hierarchy of evidentials?	*z*	0.800	*z*	0.625
ARG	Recipient arguments are marked	*d*	0.800	*d*	0.500
ARG	Passive constructions are marked	*d*	0.800	*d*	0.625
ARG	Obliques in transitive constructions	*d*	0.800	*d*	0.500
NP	What is the morphological composition of adverbial demonstratives (as compared to adnominal dem.)?	*b*	0.800	*b*	0.500
ARG	Obliques in intransitive constructions	*b*	0.800	*b*	0.500

		{Baure, Itonama, Kanoé, Movima, Mekens}		{Baure, Itonama, Kanoé, Karo, Kwazá, Lakondê, Mekens, Wari'}	
ARG	Which independent pronominal constituent is morphosyntactically treated as S:	*a*	0.800	*a*	0.750
ARG	Reciprocals are marked by	*a*	0.800	*a*	0.875
NP	In possessive constructions with pronominal possessor, is the POSSESSOR usually marked? (state most frequent construction; NB: if marking depends on alienability, state marking of possessor in ALIENABLE constructions)	*a*	0.800	*a*	0.625
NP	In possessive constructions with nominal possessor, is the POSSESSOR usually marked? (state most frequent construction; NB: if marking depends on alienability, state marking of possessor in ALIENABLE constructions)	*a*	0.800	*a*	0.625
NP	Are demonstratives related to third person pronouns?	*a*	0.800	*a*	0.750
ARG	Which constituent full NPs are morphosyntactically treated the same in basic constructions:	*a*	0.800	*a*	0.750
SUB	Temporal clauses that can mark subject agreement	*Y*	0.800	*Y*	0.625
SUB	Temporal clauses with bound subordinator	*Y*	0.800	*Y*	0.625
TAME	Is realis mood marked morphologically?	*N*	0.800	*N*	0.875
TAME	Is perfective marked morphologically?	*N*	0.800	*N*	0.875
TAME	Is perfect marked morphologically?	*N*	0.800	*N*	0.750
TAME	Is imperative for 1st person only marked morphologically?	*N*	0.800	*N*	0.750
TAME	Is dubitative marked morphologically?	*N*	0.800	*N*	0.750
NP	How many distance contrasts do adverbial demonstratives encode?	*2*	0.800	*2*	0.500