Differentiating 451 languages in terms of their segment inventories

Vladimir Pericliev and Raúl E. Valdés-Pérez*

Mathematical Linguistics Department,
Institute of Mathematics & Informatics, bl.8,
Bulgarian Academy of Sciences, 1113 Sofia, Bulgaria
peri@math.bas.bg

Computer Science Department,
Carnegie Mellon University,
Pittsburgh, PA 15213, USA
valdes@cs.cmu.edu

Abstract

A "segment niche" for a language is a subset of its segment inventory that can distinguish this language from all other languages. This paper gives the computer-generated segment niches for the 451 languages in the UCLA Phonological Segment Inventory Database (UPSID-451). It is shown that languages usually possess a multitude of segment niches, but not all of these alternatives are equally representative of a language's idiosyncrasy. A criterion is suggested for choosing among these alternatives: Preferring the niche for a language which contains segments with the smallest frequencies of occurrence in other languages (this niche, in effect, can be considered as more "typical" for the language). Some observations are offered regarding the size and structure of the computed "typical" niches.

1. Introduction

Two basic and interrelated goals of linguistics are to describe the similarities and the differences between the languages of the world. The achievement of significant results in both trends (often referred to as the "generalizing" and the "individualizing" approach respectively, e.g.. Greenberg, 1973: 163; Croft, 1990: 43) depends to a great extent on the archiving of data. Ferguson, for instance, writes (1978: 26) in this context "The need is urgent for reliable, detailed, comparable cross-linguistic data, accessible to researchers by topic". Notable success has been achieved in this direction, as e.g. the Stanford Phonological Archive, the UCLA Phonological Segment Inventory Database in phonetics-phonology; Greenberg, 1966, Hawkins, 1983, Dryer, 1992 in word order; Murdock, 1970 in kinship terminology, to mention but a few of the familiar databases in some areas. A complementary step is the automated discovery of new and interesting knowledge in the large bulks of data already available, because such searches would normally be computationally hard, and hence beyond human reach. Linguistic databases are indeed built with this aim in mind. We have been involved in the development of sophisticated computational tools for knowledge discovery in science, and linguistics in particular, modeling the individualization task (a program for multiple classes discrimination and its linguistic applications is described in Pericliev & Valdés-Pérez (1998a, 1998b, 1998c); computer science oriented descriptions are Valdés-Pérez & Pericliev 1997, Valdés-Pérez, Pereira & Pericliev 2000) or the generalization task (for a program discovering statistically significant universals, cf. Valdés-Pérez & Pericliev 1999). For a survey of some linguistic discovery systems, cf. Pericliev (1999).

In this paper, we address the task of language individualization, computing the differences between the world languages in terms of the segment inventories these languages employ. The distinctive set of segments a specific language possesses shows the position this language occupies in the space of all segments, so can be appropriately called a "segment niche" for that language. Traditional descriptive grammars of a specific language often deal with the sound idiosyncrasies this language has, contrasting it from other closely genetically/areally related languages. This question however can be dealt with in a wider linguistic context, the segment niche of a language serving as a phonetic-phonological definition of this language, and placing it in a unique position among the world's languages. In the paper, we list the segment niches for the 451 languages whose segment inventories are included in the UPSID database (Maddieson 1984; Maddieson & Precoda 1991). We investigate some aspects of the structure of the identifying segment niches, show that there is generally a large number of alternative niches, and propose an intuitive method for choosing the more "typical" one(s) for a language. Methodologically, our study illustrates the possibilities for a purely computational approach to some aspects of typology. In practical terms, in addition to serving as phonetic-phonological definitions of languages, our results may form the basis for automatic language recognition, a task recently discussed by Hombert & Maddieson (1999).

The paper is organized as follows. Section 2 briefly describes the UPSID database and our computational tools. Section 3 is an overview of the results of our computations, and the next sections present a discussion of these results. First, some observations are given concerning the size and structure of segment niches (Section 4). Then, the question of alternative segment niches is treated, showing that not all alternatives are equally representative of a language's idiosyncrasy, and proposing a criterion for choosing "typical" niche(s) for a language (Section 5). In Section 6 we show that in languages with alternative 1-segment niches (i.e. in languages possessing idiosyncratic segments) there exists a strong feature relatedness between the idiosyncratic segments of a language. Finally, in Section 7, we summarize the results, noting that a niche may be interpreted as a nearly exceptionless universal, prohibiting segment co-occurrence, which is violated by only one language of the database, viz. by the language this niche serves to discriminate. The Appendix lists the "typical" segment niches for the languages studied.

2. The data and the programs

We base our investigation on the most detailed collection of segment inventories of the world languages, compiled by Maddieson and colleagues at UCLA (Maddieson 1984). Originally, the *UCLA Phonological Segment Inventory Database* (UPSID) consisted of 371 languages. A later corrected and expanded version (Maddieson & Precoda 1991) contains already a 451 language sample, and we use this later version, known as UPSID-451, for our computations.

UPSID-451 contains phonologically contrastive segments of the world languages. For each language in the database, a segment inventory is included containing segments that have lexically contrastive function. The features assigned to a segment exhibit its most characteristic phonetic form. They all are (with rare exceptions) positively specified. The languages included are chosen on a quota principle, to the effect that all major language families are covered, but only one language from any small grouping is included. The language chosen to cover a small grouping would as a rule be described in trustworthy linguistic sources.

Our goal was to compute how the world languages contrast in terms of their segment inventories. In particular, we wanted to find the segment niches for the languages in the database, whereby by a segment niche for a language (or niche for short), we understand the collection of one or more segments from its segment inventory that distinguishes this language from all other languages.

In general, a language L1 may be said to differ from the remaining languages L2...Ln in terms of their segment inventories in two distinct ways. First, L1 may differ from L2...Ln by a set consisting of only positively specified segments actually possessed by L1 (a positive niche). Secondly, L1 may differ from L2...Ln by a set comprising some negatively specified, or segment(s) absent from L1 (a negative niche). Thus, for instance, one (of many) positive niches for Hungarian is [øː, ø], signifying that Hungarian possesses a long mid front rounded vowel and a mid front rounded vowel, which taken together are unique to this language; the negative niche [øː, not-ʏ], states that Hungarian possesses the first segment, a long mid front rounded vowel, but lacks the second, a lowered high front rounded vowel, and these circumstances distinguish it from the remaining languages.

Clearly, the use of negative niches allows the discrimination of some languages that may not possess positive niches. A case in point is the situation when a language has a segment inventory that is a proper subset of the segment inventory of another language. The language with the subsumed inventory cannot have a positive niche because it cannot possess a segment absent in the other language. The absence of a segment, which is present in the language with subsuming inventory, will allow formulating a negative niche.

We have developed techniques for discovery of both types of differences, but in the present study stick to positive niches, using negative only when the positive ones fail to provide discrimination. We believe that for the task at hand it is more intuitive to describe differences between languages in terms of segments the languages do have rather than in terms of segments they do not have.

Two specifics in the design of UPSID-451, viz. the presence of ambiguous and anomalous segments in inventories, are of direct relevance as to whether two languages contrast or not, so we turn to how we handle these problems.

In the cases when there is insufficient information about some specific segment in some language UPSID-451 will use this segment ambiguously. For instance, the segment [”tʲ”] in UPSID-451 denotes EITHER a palatalized dental plosive OR a palatalized alveolar plosive and the presence of this segment in the inventory of some language implies that it is unknown to the database whether the segment is actually a palatalized dental plosive (notated [t̪ʲ] in UPSID-451) or a palatalized alveolar plosive (notated [tʲ]). The consequence of this is that the superficially different pairs of segments [”tʲ”]--[t̪ʲ] and [”tʲ”]--[tʲ] cannot reliably contrast. Accordingly, our programs forbid differentiations between two languages based on such UNCONTRASTABLE (pairs of) segments. It may be noted that the ambiguous segment [”tʲ”] is but one example of the large class of consonants for which it is not specified in the descriptive sources whether they are dental or alveolar (and there are further ambiguous codings in UPSID-451 which our programs take into consideration).

Besides ambiguous segments, the segment inventories of languages in UPSID-451 may include a set (possibly empty) of segments which can be considered anomalous in a language. These are segments "with a somewhat doubtful or marginal status in an inventory, but with sufficient claim to be included that they were not simply eliminated from consideration" (Maddieson 1984: 170). A segment is classed as “anomalous” if any of the following conditions holds: The segment is (1) rare (extremely low lexical frequency), (2) loan (occurs in unassimilated loans), (3) abstract (posited underlying segment), (4) derived (segment possibly derivable from others), or (5) obscure (occurring in a particularly vague or contradictory description). To make the generated positive niches plausible and reliable, we stipulate that niches must not contain an anomalous segment (to avoid claiming that a language L1 discriminates from another language L2 by a segment which is doubtful or marginal for L1). So for constructing a niche for L1 we choose segments from the REDUCED SEGMENT INVENTORY of L1, i.e. the inventory resulting from removing all anomalous segments from it. And, secondly, to ensure that L2 indeed does not possess the segment included in the niche of L1, and used to contrast L1 from L2, we check that this segment is reliably absent from L2, which requires this check to be made in the NON-REDUCED SEGMENT INVENTORY of L2 (i.e. the inventory containing all segments, the anomalous ones including).

It should be noted that our conservative policy in discrimination---both as regards the treatment of ambiguous and of anomalous segments---ensures maximal reliability of the resultant positive niches.

In broad outline, the algorithm of the program for discovery of positive niches is the following. The program selects a language from the database and inspects whether anyone of its (non-anomalous) segments is absent from the (non-reduced) segment inventories of all other languages. If this is the case, the program finds all other idiosyncratic segments and the result is reported as (alternative) 1-segment niches for that language. If not, the program hypothesizes 2-segment niches (by forming all unordered pairs of segments from the language's (reduced) segment inventory) and checks which of these hypotheses hold in the database. This failing, the program consecutively hypothesizes and tests more complex niches of size 3, then of size 4, then of size 5 etc. until a solution is found (or a failure is reported on exhausting all hypotheses).

It will be clear from this outline of the basic algorithm, that our tools find the MINIMAL (=shortest) niches, individuating the languages, as well as all their ALTERNATIVES of the same length.

The search for negative niches is carried out by a rather complicated algorithm that compares the target language against each other language, then minimizes the segments that are needed to differentiate the target from every other language. These comparisons are not limited to segments that the target language possesses, which is why the niches can turn out negative. Anomalous segments are treated as "missing values", which means that two languages are never contrasted by a segment if either language has an anomalous value for that segment. This conservative procedure guarantees stable niches: resolving an anomalous value either way (as truly belonging to the language or as extraneous) will not change the accuracy of the niches we have found.

It should be noted that the UPSID-451 database is designed to support machine investigations, and Maddieson & Precoda (1991) do provide some useful programming facilities (UPSID and PHONEME) that can be used as a quick reference to find occurrences of some particular segment or to look up the segment inventory of a particular language. These facilities however cannot be used for more complex computations, as required by our task of segment inventories contrasts.

3. Overview of the results

We ran our programs on languages in the UPSID-451 database. The Appendix lists the computer-generated niches for these languages, starting with the languages using positive niches, in ascending order of niche size. One can read off from the Appendix that, for example, Fijian has the 1-segment niche [^{n}r] (a prenasalized voiced alveolar trill), which differentiates it from all other languages in UPSID-451, that for Papago this is the 2-segment niche, [d.,n̲] (a voiced retroflex plosive and a voiced palato-alveolar nasal), that Tarok needs a 3-segment niche ["d", ʒ, k͡p] (a voiced dental/alveolar implosive and a voiced palato-alveolar sibilant fricative and a voiceless labial-velar plosive), and so on.

Two languages, Dyirbal and Yidiny (both belonging to the Australian family, Pama-Nyungan subfamily) are in principle indistinguishable from one another since they have identical (reduced) segment inventories. They have a shared niche, making them distinct from the rest of the languages, given at the end of the Appendix. 25 further languages cannot be distinguished---employing only positive niches---from some other language(s). Below, we show the twenty five languages that fail discrimination in the left-hand column, and the language(s) with which they fail to discriminate in the right-hand column.

Languages failing discrimination	Languages causing the failure of discrimination
Ainu	Sa'ban, Cayuvava
Chuave	Kanuri, Luo, Sandawe, Camsa, Gwari, Yaqui
Dera	Dagbani, Tarok, Sama
Birom	Tampulma
Bisa	Amo, Tampulma
Efik	Aghem, Amo, Bete, Lelemi, Tampulma

Western Desert	Arrernte
Piraha	Jingpho, Kurukh, Sama, Cofan, Huari, Chatino
Yaqui	Kanuri, Luo, Lua
Bandjalang	Ivatan
Songhai	Burushaski, Bambara, Kolokuma, Tampulma, Tarok, Tera
Iwam	Tarok, Iban
Koiari	Dahalo
Nasioi	Jingpho, South Kiwai, Bruu, Nyah, Itonama, Paya, Ticuna, Chatino, Kawaiisu, Kefa
Rotokas	Dahalo
Suena	Jingpho
Taoripi	Saami, Burushaski, Tarok, Gelao, Lai, Tetun, Amharic
Nera	Burushaski, Tarok
Nubian	Saami
Batak	Jingpho,Burushaski, Tarok, Sama
Hawaiian	Jingpho, Dani, Kurukh, Sama, Sui, Tetun, Itonama, Paya, Chatino, Kawaiisu, Tol, Kefa
Iban	Sama
Roro	Itonama, Paya
Campa	Amuzgo
Moxo	Amuzgo, Dahalo

The first ten languages above fail discrimination since their (reduced) segment inventories are subsumed by the inventories of the languages in the right-hand column, whereas the remaining fifteen languages happen not to have a non-anomalous segment contrastable with the segments of some other languages. The Appendix lists the negative niches that these languages need.

Table 1 presents some statistics on positive niches, concerning their size (column 1), the number of languages using a niche of this size (column 2), the average sizes of the reduced segment inventories of these languages (column 3), and the corresponding average number of alternative niches (column 4). Thus, row 2 e.g. indicates that 213 languages, whose mean reduced segment inventory size is 29.5, need at least the simultaneous presence of 2 segments to put them apart from the rest of the languages, and that, one the average, there are 19 alternative 2-segment niches per language.

=================

Table 1 goes about here

=================

We note that not all alternative niches, as calculated in Table 1, are actually given in the Appendix, for reasons to become clear in Section 5.

The next sections comment on the results in the Appendix, as summarized in Table 1.

4. Some observations on the size and structure of niches

Table 1 (column 1) shows that positive niches vary in size from 1 to 7 segments.[1] From columns 1 and 2 it is seen that about 3/4 of the languages from the database have niches of one or two segments. Most numerous (viz. 213) are the languages needing 2-segment niches.

Comparing the sizes of the positive niches (hereafter just "niches") and the average sizes of the corresponding segment inventories (columns 1 and 3), we see a (negative) correlation between these two parameters, viz. the smaller the niches able to distinguish some languages, the larger their average segment inventories, and the larger the niches, the smaller the inventories. This correlation is very slightly violated only in the case of languages needing 6-segment niches (row 6 of Table 1). The correlation could be attributed to the interaction of the following two circumstances. The first one (noted by Maddieson 1984: 10) is that a smaller inventory has a greater probability of including a given common (frequent) segment than a larger one, and a larger segment inventory has a greater probability of including an unusual (infrequent) segment type than a smaller one. The second circumstance is that a segment set of a fixed cardinality (=size) is more likely to distinguish a given language if it consisted of unusual segments rather than if it consisted of more common segments.[2] Therefore the languages with larger segment inventories, which will presumably contain unusual segments, will use for discrimination a smaller number of these than the languages with the smaller segment inventories, containing exclusively more common segments. On the other hand, a negative correlation between niche size and the number of segments can be expected even if all segments occurred with equal frequency.

Do any patterns emerge regarding the structure of niches? In the first place, we examined the distribution of vowels and consonants in niches, grouping under the class of vowels all segments marked as (simple) vowels and diphthongs in UPSID-451, and under the class of consonants all remaining segments. The results are summarized in Table 2.

=================
Table 2 goes about here
=================

Table 2 reveals patterns like the following: If a language uses a 1-segment niche, then it is more likely that it consists of a single consonant rather than a single vowel (240 out of 342 total 1-segment niches or 70.1%); or within niches of equal size, the most frequently occurring niches are those in which consonants outnumber vowels, etc. The general conclusion to be drawn from the table is that consonants are more frequently used in niches than vowels, a fact that can simply be attributed to their higher overall frequency of occurrence within the segment inventories in UPSID (Maddieson 1984: 9).

In the second place, we looked for other patterns in the niches. As mentioned above, niches will tend to comprise rarer segments, since these will generally have a higher discriminative power and we explored this line of research. At present, we dispose of no definitive answer to the question of what segments are rare in the world's languages, and it is likely that no general answer can be given to this question (for a general discussion, and a discussion of cross-linguistic frequency of some specific types of sounds, cf. e.g. Maddieson 1984; Lindblom & Maddieson 1988; Laver 1991). For the purposes of our study, we investigated the occurrence in segments of features singled out under the rubrics "secondary articulations or accompanying features" and "phonation and other types" by Maddieson & Precoda (1991). In particular, these included the features labialized, palatalized, velarized, pharyngealized, nasalized, nasal release, prestopped and lateral release (under the first rubric), and the features voiceless, voiced, aspirated, laryngealized, long, breathy, overshort and preaspirated (under the second rubric). Most of these features signify a certain articulatory complexity of segments in comparison with the segments' "basic" (non-labialized, non-palatalized, etc.) types, and can hence be expected to be less frequent cross-linguistically.

There are 676 (positive) niches in the Appendix, comprising a total of 1148 segment tokens (occurrences). Of these, 985 segments or 85.8% contain at least one feature of the above list. The number of features from this list these segments use is 1577, or about 1.6 features per segment. The segments occurring in niches therefore may be said to typically comprise features indicating secondary articulation, accompanying features or different phonation types.

In the third place, we investigated the least frequent segment in each niche, or what is the same, the segment with the highest discriminative power. 1-segment niches, of course, will contain only (non-anomalous and contrastable) segments occurring in a single language, so we addressed larger niches. It was found that the most infrequent segments in 2-segment niches fall in the range 1-35 segment occurrences in UPSID-451, 3-segment niches fall in the range 1-60 segment occurrences,[3] 4-segment niches in the range 5-66, 5-segment niches in the range 33-61, 6-segment niches in the range 33-74, and 7-segment niches in the range 50-50 (the Appendix contains only one 7-segment niche, for Tetun). These figures are yet another indication of the correlation between a niche's size and the frequency of its constituent segments. At the same time, the overlap of the ranges of pairs of consecutively larger niches (disregarding only niches of size=1) shows the more complex interaction between segments, as they jointly act as discriminants, a fact we referred to in Note 2. Thus, it may on occasion happen that the rarest segment in a niche of size=2 is in fact more frequent than the rarest segment in a niche of size=3, 4 or 5; the rarest segment in a niche of size=3 is more infrequent than the rarest segment in a niche of size=4 or 5, etc.

Finally, we investigated whether the niches’ constituent segments share some features that bind these segments together (making sense only in niches larger than 1-segment). Of course, those niches from Table 2 that consist of vowel segments only, or of consonantal segments only, are exemplary, since they share the feature vowel or consonant, respectively. However, there exist further possibilities for feature sharing in segments belonging to the same niche, and we explore these next.

We found that out of 334 niches that possess more than one segment, 193 (i.e. 57.7%) have component segments any one of which shares *at least* one feature with all other component segments. The shared features may significantly exceed one, as in most of the following examples:

Hadza

2-segment niche: [ǁʰ, ŋ̥ǁʔ]

Segment feature description: [ǁʰ]-[voiceless aspirated alveolar lateral affricated click]

Segment feature description: [ŋ̥ǁʔ]-[glottalized nasalized voiceless alveolar lateral affricated click]
Shared features (5): [voiceless alveolar lateral affricated click]

Andoke
2-segment niche: [ã̟, ɑ̃]

Segment feature description: [ã̟]-[nasalized low front unrounded vowel]

Segment feature description: [ɑ̃]-[nasalized low back unrounded vowel]
Shared features (4): [nasalized low unrounded vowel]

Mazatec
2-segment niche: [ⁿ·d.z., ⁿdz]
Segment feature description: [ⁿ·d.z.]-[prenasalized voiced retroflex sibilant affricate]
Segment feature description: [ⁿdz]-[prenasalized voiced alveolar sibilant affricate]
Shared features (4): [prenasalized voiced sibilant affricate]

Auca
3-segment niche: ["ẽ", æ̃, ĩ]
Segment feature description: ["ẽ"]-[nasalized mid front unrounded vowel]
Segment feature description: [æ̃]-[nasalized raised low front unrounded vowel]
Segment feature description: [ĩ]-[nasalized high front unrounded vowel]
Shared features (4): [nasalized front unrounded vowel]

Kwaio
3-segment niche: [ŋʷ, ᵑgʷ, xʷ]
Segment feature description: [ŋʷ]-[labialized voiced velar nasal]
Segment feature description: [ᵑgʷ]-[prenasalized labialized voiced velar plosive]
Segment feature description: [xʷ]-[labialized voiceless velar fricative]
Shared features (3): [labialized velar consonant]

The common features in all the segments of a niche allow characterizing the niche. For example, we may say that Kwaio differs from all other languages (in UPSID-451) by a combination of 3 of its labialized velar consonants, Auca by a combination of 3 of its nasalized front unrounded vowels, etc. Thus, our counts show integrity of niches that seemingly exceeds randomness. Thus, niches can be described in more general, feature-centered ways.

5. Alternative niches and how to choose among them

Languages discriminate from one another in a great number of ways, as shown by column 4 of Table 1, which gives the average number of alternative niches of different sizes. These averages, however, conceal much variation.

Let us look at the alternatives within 1-segment niches, i.e. cases where each of many single segments of a language can discriminate it from the other languages. Here, 71 languages do not actually have alternatives (i.e. possess only one idiosyncratic segment), while most of the other 53 languages have only a couple of such alternatives. Several languages however have a significant number of alternatives relative to their total (reduced) segment inventories. The leader is !Xóõ with 65 alternatives, which amounts to 46% of its total (reduced) segment inventory of 140 segments. There follow several languages with considerably less alternatives (from 7 to 19), but whose idiosyncrasy is still impressive in that the unique segments they possess form a large part of their total sound inventories: Parauk (19/77 24%), Shilha (7/31 22%), Kabardian (8/55 14%) and Hmong (8/55 14%), Archi (10/75 13%), and Irish (7/63 11%). Though the Caucasian languages Kabardian and Archi are believed by some linguists to belong to the same language family, it is seen that the large proportion of idiosyncratic segments in a language is not correlated with its genetic origin, as the other languages above belong to widely diverse language families. It seems to be correlated (as could be expected from our discussion in the previous section) with the largeness of inventories, but is not entirely limited to such languages, as Shilha shows, which has the average of 31 segments for this corpus.

From the 213 languages with two-segment niches only 37 do not have some alternative. Here the mean number of alternatives drastically increases (more than six times in comparison with 1-segment niches), and one language, Hadza, has 320 alternative idiosyncratic segment pairs. Those exceeding 100 alternatives, in diminishing order, are Sedang (306), Lithuanian (208), Mazahua (165), Hindi-Urdu (137), and Kurdish (136).

From the 57 languages with distinctive triplets, 10 do not have alternatives, their mean number now being 29.9 triplets per language. Bella has 220 alternatives. Ormuri with 120 and Lushootseed with 115 are the only other languages in which this number exceeds 100.

From the 16 languages with distinctive 4-tuples, 2 have a single niche. The mean number of alternatives increases to 30.1, and the language with a maximum of alternatives, viz. 228, is Quileute. No other languages exceeds 100 alternatives.

Only 14 languages need 5-segment, 6-segment or 7-segment niches to be distinguished from the other languages and they generally have a smaller number of alternatives (Sa'ban is an exception with its 43 alternative quintuples, and so is Tagalog with 40 alternative sextuples).

The existence of this multitude of ways to differentiate a language from the others creates some problems. In the first place, we would like to limit these possibilities, especially if we want to view niches as phonetic-phonological definitions of languages (it would be simplest to have just one definition of one and the same object). In the second place, and more importantly, our treatment of alternative niches so far implies that all of them are EQUALLY representative of a language's idiosyncrasy, which turns out not to be the case. An example will make this point clearer.

Tigre, for instance, has the following 8 alternative niches:

[a̟ː”ə“]　　[a̟ːdʒ]　　[a̟ːg]　　[a̟ːk′]
[a̟ːtʃ]　　[a̟ːtʃ′]　　[a̟ːts′]　　[a̟ːz]

anyone of which is perfectly legitimate in the sense that it suffices to discriminate this language from all other languages. On a closer look, however, it may be seen that not all of these niches represent equally well the specificity of Tigre, or its sound traits that put it apart from all the other languages of the world. Thus, the segment [a̟ː] recurs in all 8 niches, so it must be a component of any niche we might choose to use, but, as far as the second component of the niche is concerned, we have a choice between 8 alternative segments, viz. [”ə“,dʒ,g,k′, tʃ, tʃ′, ts′, z]. Some of these Tigre segments occur infrequently in other languages, hence they can be considered more "typical" of Tigre than the more widespread segments. Looking at their frequencies of occurrence in UPSID-451 (abstracting from their occurrence in Tigre itself), we get the following picture: [”ə“] (72 occurrences), [dʒ] (107), [g] (235), [k′] (62), [tʃ] (181), [tʃ′] (43), [ts′] (25), and [z] (58). Clearly, then, if we are after typicality, we should choose the most infrequent segment in the list, [a̟ː],

occurring in only 5 other languages, which together with [tsʼ] (occurring in only 26 other languages), would yield the most "typical" niche for Tigre. More generally, a TYPICAL niche for a language will comprise the more unusual segments occurring in that language in comparison with the other, less typical, alternatives. We thus come up with the following simple intuitive criterion for choosing a TYPICAL NICHE from all alternative niches:

> *From a set of alternative niches for a given language, choose the niche(s) containing the segments with the lowest frequency in the database*.

This criterion was implemented in our system by counting the occurrences in the segments database, summing these counts for each niche, and then choosing the niche with the smallest count. The results in the Appendix reflect this criterion and list only typical niches for the languages. After applying this filter some alternative typical niches may still remain. Of course, this will be the case in all 1-segment niches, since all the alternatives for one language will consist of a segment with zero frequency of occurrence in the remaining languages. So would alternatives with equal total frequencies in niches larger than 1-segment. But these are not unwelcome consequences insofar as we would as a rule be interested in showing all segments occurring in just one language and, more generally, have all equally intuitive larger niches, from which we could choose randomly, if need be (e.g. for providing a phonetic-phonological definition of some language).

6. Feature relatedness between alternative 1-segment niches

1-segment niches are of considerable interest since they pertain to segments unique to a particular language. We posed the question: If one language possesses several idiosyncratic segments, are these segments related in terms of their feature specifications? In other words, do these segments share some features that bind them together?

There are several cases of languages possessing idiosyncratic segments, which we present below, accompanying each with an example:

Case 1: All unique segments share some feature(s)
Akan
[çʷ]-[labialized voiceless palatal fricative]
[cçʷ]-[labialized voiceless palatal affricate]
[ɟʝʷ]-[labialized voiced palatal affricate]
[ɲʷ]-[labialized voiced palatal nasal]
Shared features (3): [labialized palatal consonant]

Case 2: A group of unique segments shares some feature(s), another group shares other feature(s)
Kashmiri
[dʒʲ]-[palatalized voiced palato-alveolar sibilant affricate]
[tʃʲ]-[palatalized voiceless palato-alveolar sibilant affricate]
Shared features (4): [palatalized palato-alveolar sibilant affricate]

[ə̃ː]-[long nasalized higher mid central unrounded vowel]

[ɨ̃ː]-[long nasalized high central unrounded vowel]
Shared features (5): [long nasalized central unrounded vowel]

Case 3: A group of unique segments share some feature(s), but an unrelated segment remains
Hmong
[ᵑkʰ]-[prenasalized voiceless aspirated velar plosive]
[ᵑq]-[prenasalized voiceless uvular plosive]
[ᵑqʰ]-[prenasalized voiceless aspirated uvular plosive]
[ⁿ.t.s.]-[prenasalized voiceless retroflex sibilant affricate]
[ⁿ.t.s.ʰ]-[prenasalized voiceless aspirated retroflex sibilant affricate]
[ⁿ̪t̪s̪ʰ]-[prenasalized voiceless aspirated dental sibilant affricate]
[ⁿtʃʰ]-[prenasalized voiceless aspirated palato_alveolar sibilant affricate]
Shared features (3) -[prenasalized voiceless consonant]

[ɛɯ]-[lower mid front unrounded to high back unrounded diphthong]
(Isolated segment)

Case 4: No group of unique segments shares any feature(s)
Somali
[d̰.]-[laryngealized voiced retroflex plosive]
(Isolated segment)
[ʉ̞]-[lowered high central rounded vowel]
(Isolated segment)

In case 1, all alternative unique segments for a language have one or more features in common. In Akan, all alternatives are grouped together by the feature set [labialized palatal consonant]. In case 2, the alternative unique segments for a language can be partitioned in two or more groups, the segments in each group being bound together by a feature set. The four alternatives of Kashmiri form two groups of segments, the palatalized palato-alveolar sibilant affricate group, consisting of two segments, and the long nasalized central unrounded vowel group, consisting also of two segments. In case 3, some of the segments are tied into groups by a set of features, but an isolated segment remains, sharing no features with the group. In Hmong, seven segments are classed together by the feature set [prenasalized voiceless consonant], while the segment [ɛɯ] remains unrelated to the rest of alternatives. Finally, in case 4, the segments exhibit no feature similarity, as in Somali which differentiates from the other languages in UPSID-451 by the consonantal segment [d̰.] and the vowel segment [ʉ̞] which are completely unrelated insofar as their feature specifications are concerned.

In each of the cases 1, 2, and 3 there exists a noticeable relatedness between the alternative idiosyncratic segments for one language, in the sense that all alternatives can be clustered into one or a couple of groups of segments sharing some feature(s), and only in case 4 is such a relatedness absent. The degree of this relatedness can further be judged by the number of shared features defining a cluster. We conducted estimates of the number of languages in the Appendix falling under each of these cases, as well as of the degree of segment relatedness in each cluster-defining feature set.

Our counts showed the following. First, from a total of 53 languages having alternative idiosyncratic segments, 49 (i.e. 92%) fall under one of the Cases 1, 2 or 3. More exactly, the distribution among the 4 cases is: Case 1=41 languages, Case 2=3 languages, Case 3=5, Case 4=4 languages. These figures indicate that, in their vast majority (92%), the corresponding languages individuate by segments which are themselves related; besides, most often all the segments group into just one cluster, defined by a set of the features these segments share. Secondly, we computed the degree of relatedness, or the strength of similarity, holding together the segments in one grouping. We found that their similarity is quite significant, amounting to an average of 2.7 shared features per grouping.

Thus, it turns out to be an interesting characteristic of languages that, if they possess several unique segments, they would tend to significantly relate to one another. Indeed, it is a remarkable fact that, for example, all the idiosyncratic segments of Shilha, which are 7 in number (see Appendix), are pharyngealized consonants; all 6 segments of Arrernte are voiced plosives, those of Rutul are pharyngealized uvular consonants, while all 6 segments of Nama are aspirated voiceless clicks; all 6 segments of South Nambiquara are laryngealized sounds (vowels or consonants). Or, that all 7 unique segments that Irish has should turn out to be consonants, or that all 10 such segments of Archi should also be consonants, but all of them voiceless. Or, further, that in languages having 2 or 3 unique segments these segments should share as much as three or four features, as in Wolof, Russian, Changzhou, Lungchow, Bruu, Neo-Aramaic, Vietnamese. The alternative unique segments a language possesses are not random segments from its inventory, but are rather ones that are strongly bound together by the features they share.

7. Conclusion

We have described how the languages of the world differ in terms of the segment inventories they employ, listing the niches for the languages contained in the UPSID-451 database. In general, there is a significant variety of niches discriminating a language from all the others, but some niche(s) can be considered more typical for a given language in that they contain segments that are less usual in the other languages, and we listed these typical niches in the Appendix. The size of (positive) niches varies from 1 to 7 segments, but by far the most common are the 1-segment and 2-segment niches, utilized by nearly ¾ of all languages. We made also some quantitative and qualitative observations on the content of the generated niches. A (positive) niche will generally tend to include more consonants than vowels, and more unusual, rather than more usual, segments, which further, share some features in common. The segments in niches tend to have features pertaining to secondary articulations or diverse phonation types. Alternative 1-segment niches for one language show significant inter-relatedness in terms of their feature specifications.

Finally, we note that, since a niche is a sequence of segments, unique to just one language, it can be viewed as a universal prohibiting the co-occurrence of this sequence of segments. This interpretation allows us to look at a niche for some language as a nearly absolute universal, valid in all but this particular language, or in 450/451 (=99,8%) of the cases. Thus, 1-segment niches of the form [A] state that segment A does not occur in any language, 2-segment niches [A, B] state that segments A and B do not co-occur, and so on for larger niches.[4] For instance, Seneca's niche [”dz”,”o”,ɛ̃] is a universal, stating that these three segments do not occur together in any language, and will be valid in all languages except Seneca. The dual interpretation of our results, as discriminant sets or nearly absolute universals, is yet another indication of the close relation between both approaches to language, the individualizing vs. the generalizing, we referred to at the beginning of this article. Our empirical study thus would be of potential interest also to investigators of phonological universals.

Appendix. The niches for the languages in UPSID-451

The Appendix includes the "typical" segment niches for the languages in UPSID-451, i.e. those niches that pass the selection criterion given in Section 5. First, the languages using positive niches are listed, followed by those needing negative niches to individuate them from the rest of the languages. Languages are given in alphabetical order, starting with the languages employing niches of minimal size.

The notation of segments below (and throughout the article) is that of Maddieson & Precoda (1991), which basically follows the IPA conventions. To alleviate reading, here we give a listing of some diacritic marks or other symbols used, some of which may NOT be standard IPA conventions (e.g. double quotes "b" denote an unspecified dental or alveolar consonant or is undifferentiated ‘mid’ for vowels). The complete notational details can be found in Maddieson & Precoda (1991).

$b^{ɣ}$	velarized
b^{w}	labialized
b^{j}	palatalized
b̰	laryngealized
bʔ	glottalized
$b^{ʕ}$	pharyngealized
b̃	nasalized (of flaps approximants, vowels)
ŋb	nasalized (of clicks)
ɡb	voiced (of clicks)
bx	velar-fricated (of clicks)
b^{h}	aspirated or breathy
b̤	with breathy release
^{h}b	preaspirated (of stops) or voiceless (of nasals, approximants, vowels, clicks)
b̪	dental
"b"	unspecified dental or alveolar for consonants; undifferentiated ‘mid’ for vowels
b.	retroflex
b<	implosive
b’	ejective
^{n}b	prenasalized (also ^{ŋ}b)
b̆	overshort (of vowels)
rr	voiced r-sound
D	voiced tap

I. Languages discriminated by positive niches

1-segment niches

ACOMA--[ɕʼ] or [”D̪̰”] or [s.ʼ]

AKAN--[çʷ] or [cçʷ] or [ɟʝʷ] or [ɲʷ]

ALAWA--[ⁿ·d.]

AO--[ɮ̦]

ARAUCANIAN--[t.θ.]

ARCHI--[”ɬʷː”] or [χˤː] or [χʷˤː] or [qχʼ] or [qχˤ] or [qχˤʼ] or [qχˤʼː] or [qχʷʼ] or [qχʷˤ] or [qχʷˤʼ]

ARRERNTE--[bᵐ] or [d.ⁿ·] or [d̪ⁿ̪] or [d̠ⁿ̠] or [dⁿ] or [gⁿ]

ASHUSLAY--[”t̰s̰”]

AVAR--[”tɬʼː”] or [”tɬː”] or [kxʼː] or [kxː] or [qχː]

BEEMBE--[pfʱ]

BRETON--[”ø”] or [ɛ̃ɔ̃] or [ɛo] or [aɛ] or [ãɔ̃]

BRUU--[ʌ̤] or [ɒ̤]

BULGARIAN--[t̪s̪ʲ]

BURMESE--[õũ]

CHANGZHOU--[”tʱ”] or [”tsʱ”]

CHIPEWYAN--[”tʰ”] or [kxʷʰ] or [t̪θ̪ʰ]

CHUVASH--[ø̆]

DAHALO--[ɗ̠]

DAN--[ɗ.]

EKARI--[gᴸ]

EPENA--[̃ɪ-]

EVEN--[ia̟]

EWE--[ɾ̠]

EWONDO--[̟i] or [u̟]

FIJIAN--[ⁿr]

FRENCH--[œ̃]

GA--[tʃʷ]

GELAO--[”ⁿts”] or [əɯ] or [ᶮcç]

HAMER--[ɐˤ]

HMONG--[ɛɯ] or [ᵑkʰ] or [ᵑq] or [ᵑqʰ] or [ⁿ·t.s.] or [ⁿ·t.s.ʰ] or [ⁿ̪t̪s̪ʰ] or [ⁿtʃʰ]

HOPI--[θ.]

HUASTECO--[ʒ̥]

IAI--[ŋ̊͡m̊] or [l̥.] or [n̥.]

IGBO--[”ɾʲ”] or [b̤ʲʰ] or [g̤ʷ]

IK--[”e̥”] or [”o”] or [ḁ]

INUIT--[fː]

IRANXE--[a͠i] or [bʷ] or [ɨ͠i]

IRISH--[βʲ] or [ϕ˞ʷ] or [b˞ʷ] or [m˞ʷ] or [n˞] or [p˞ʷʰ] or [ꜰ]

ITELMEN--[ɮʲ]

JACALTEC--[q˂]

JAPANESE--[ɴ]

JAVANESE--[tsʱ]

KABARDIAN--[θ̠] or [ð̠] or [ʔʷ] or [fʼ] or [ɣʲ] or [ɬʲʼ] or [ɮʲ] or [xʲ]

KAROK--[ʊː]

KASHMIRI--[ə̃ː] or [dʒʲ] or [ɨ̃ː] or [tʃʲ]

KHALKHA--[ʊi]

KHANTY--[”ĕ”] or [”o”] or [ɒ̆]

KHMER--[eə] or [ɯː]

KHMU?--[ɨa]

KIOWA--[a̟i] or [a̟͠i] or [ᵈ̪l̪]

KOHUMONO--[k͡pʰ]

KOLOKUMA--[g͡bʷ] or [k͡pʷ]

KONYAGI--[ᵑkʷ] or [ᶮc]

KOREAN--[s̪̰]

KOTOKO--[pfʼ]

KWOMA--[ϕʷ]

LAI--[”ɬʲ”] or [”ⁿdʲ”] or [”tʲʰ”] or [ᵐbʲ]

LAK--[”øˤ”] or [”tsʷː”] or [qː] or [qʷː] or [tʃʷː] or [xʷː]

LAKKIA--[ŋ̊ʲ] or [˜iẽ]

LUNGCHOW--[ĭ] or [ʊ̆] or [ɯ̯̆]

LUO--[d̪ð̪]

MALAGASY--[d.r̘ʔ] or [t.r̘ʔ]

MALAKMALAK--[ø̠]

MANDARIN--[cçʰ]

MBUM--[v̝]

NAHUATL--[ɸ̞]

NAMA--[!xʰ] or [‖xʰ] or [ǀx̊ʰ] or [ŋ̊!ʰ] or [ŋ̊‖ʰ] or [|xʰ]

NAMBAKAENGO--[”ⁿdʷ”]

NAXI--[ᶮɟʑ]

NENETS--[ð̪ʲ]

NEO-ARAMAIC--[iˤː] or [uˤː]

NEWARI--[”d̤z̤”]

NGANASAN--[kx]

NGIZIM--[ɟ]

NIVKH--[ɪe] or [t̠ʰ]

NORWEGIAN--[ɒ̝] or [æʉ] or [øy] or [ʉː]

NYAH--[i̤ə̤] or [ṳə̤] or [ɯ̤ə̤]

OCAINA--[ɲː]

PACOH--[d̠̰] or [ɨə]

PAEZ--[ɻʲ]

PARAUK--[ŋ̤] or [ɔ̤i̤] or [ʒ̤] or [a̤i̤] or [a̤ṳ] or [a̤ɯ̤] or [iɛ] or [i̤ɛ̤] or [̤ia̤] or [io] or [̤iṳ] or [l̠̤] or [n̠̤] or [ɤi] or [ɤ̤i̤] or [o̤i̤] or [ṳa̤] or [ṳi̤] or [ɯ̤i̤]

PHLONG--[vʷ]

ROMANIAN--[ea]

RUSSIAN--[ʆ] or [ʓ]

RUTUL--[ɢˤ] or [ɢʷˤ] or [qˤ′] or [qˤʰ] or [qʷˤ′] or [qʷˤʰ]

SAAMI--[”dzʲ”] or [ß̯ʲ] or [hʲ]

SAMA--[”ⁿs”]

SANDAWE--[!ʔ] or [‖ʔ] or [g!] or [g‖] or [|ʔ]

SEBEI--[ʉ̆]

SELEPET--[ɑ̝]

SELKUP--[”ɤː”]

SENADI--[”zʲ”]

SHILHA--[”dˤ”] or [”lˤ”] or [”rˤ”] or [”sˤ”] or [”tˤ”] or [”zˤ”] or [kˤ]

SHUSWAP--[”t̰ɬ̰”] or [ʁ̞ʷ] or [ʁ̞̰ʷ] or [q̰] or [q̰ʷ]

SIONA--[t̰.]

SOMALI--[ᵿ] or [d̰.]

SOUTH_NAMBIQUARA--[”ḛ̃”] or [”o”] or [”õ̰”] or [ã̰] or [ḭ̃] or [ɻ̰.]

SUI--[ɣ̰]

TACANA--[ɾ̝]

TAMANG--[tθʰ]

TARASCAN--[ɨʵ]

TEHUELCHE--[cɕ′]

TEKE--[ɱ]

TELUGU--[ß̯ː]

TERA--[ɓʲ] or [ᶮɟʝ]

TLINGIT--[χ′] or [χʷ′] or [xʷ′]

UZBEK--[ß̯̃]

VANIMO--[ɵ̃]

VIETNAMESE--[iʌ] or [uʌ] or [ɯʌ]

WAHGI--[ʟ̝̊]

WANTOAT--[”ⁿz”]

WAPISHANA--[z̰.]

WARAY--[cː]

WICHITA--[”nː”] or [ɒː]

WOISIKA--[”e̠”] or [”o̟”]

WOLOF--[bː] or [ʄː] or [l̪ː]

!XÓŨ--[”oˤː”] or [”õˤ”] or [ǁ̟] or [ǁ̟ⁿ] or [ǁ̟x] or [ǂ] or [ǂʰ] or [ǂx] or [ǀʰ] or [ǀx] or [ŋˤ] or [ɔ̃ˤː] or [ãˤ] or [ãˤː] or [aeˤ] or [ãẽˤ] or [aoˤ] or [ãõˤ] or [b′] or [d′] or [đ] or [dʒ′] or [đʒ] or [dz′] or [đƶ] or [ẽũˤ] or [gǁ̟] or [gǁ̟x] or [gǁ̟x ʔ] or [g′] or [gǀ] or [gǂ] or [g̤ǂ] or [gǂx] or [gǂxʔ] or [g̤ǀ] or [gǀx] or [gǀxʔ] or [g̤|] or [g|x] or [g|xʔ] or [ŋ̊ǁ̟ʔ] or [ŋ̊ǁ̟ⁿ] or [ŋǁ̟xʔ] or [ŋ̊ǂʔ] or [ŋ̊ǂʰ] or [ŋ̊ǂxʔ] or [ŋǀxʔ] or [ŋ|xʔ] or [ŋ‖] or [ŋ̤ǁ̟] or [ŋǂ] or [ŋ̤ǂ] or [ŋ̊ǀ] or [ŋ̤|] or [oaˤ] or [õãˤ] or [õã] or [oe] or [oiˤ] or [õĩˤ] or [ŧ] or [ŧʃ] or [ŧs] or [|x]

YAGARIA--[ʟ]

YAKUT--[yø]

YANYUWA--[ᵑkʲ] or [ⁿ̠t̠]

YUCHI--[ɸ′]

YUKAGHIR--[”ꞙ”]

ZOQUE--[”ɣ̟̃”]

ZULU--[ǀˢ] or [kʟ̥′]

2-segment niches

ACHE--[”ɹ”, ”ⁿd”]

ADZERA--[”dz”, ”oː”] or [”dz”,ɰ]

AGHEM--[ɒ, bv]

AHTNA--[dɮ, ɮ]

AIZI--[ɰ̞, k͡p]

AKAWAIO--[”z”, ɔi]

ALAMBLAK--[D, tʰ]

ALBANIAN--[ɟʄ, ł]

ALEUT--[ð̪, j̊]

ALLADIAN--[ɟ, ɥ]

AMAHUACA--[ɪ̃, ɯ̃]

AMELE--[g͡b, ʄ]

AMHARIC--[”sʷ”, hʷ]

AMUESHA--[lʲ, t.s.ʰ]

AMUZGO--[ɒ̃, kʲ]

ANDOKE--[ã̟, ɑ̃]

ANGAATIHA--[”ə̆”, ɹ]

ANGAS--[bʲ, ɨː]

APINAYE--[ʌ̃, ᶮɟ]

ARABIC--[æː, s̪ˤ]

ARMENIAN--[ð., ʁ̞]

ATAYAL--[”ts”, ħ]

AWIYA--[ɢʷ, ŋʷ]

AZANDE--[”ɜ̃”, ɹ]

AZERBAIJANI--[ʏ, ɣ̟]

BAI--[ɞ, ɯ̝]

BAINING--[ɛi, ai]

BARDI--[aː, l̠]

BASHKIR--[”ɤ”, θ̪]

BASQUE--[ɟ, t.s.]

BATS--[ʡ, ɑ̃]

BEJA--[gʷ, t.]

BENGALI--[æ̃, d̪̤]

BERTA--[”ⁿd”, ”s′”]

BETE--[ᵻ, ɰ]

BOBO-FING--[k͡p, tʰ]

BRAHUI--[”ɬ”, eː]

BRAO--[cç, ⁿtʃ]

BRIBRI--[”ẽ”, ”z”]

BURUSHASKI--[ð., d.s.]

CADDO--[”ts′”, ʊ]

CAMSA--[ɨ, t.s.]

CAYAPA--[d̠, l̠]

CAYUVAVA--[æ̃, ɨ̃]

CHAM--[cʰ, d̰]

CHAMORRO--[dz, ɹ.]

CHUKCHI--[”ə̆”, ð.]

COFAN--[ɰ, ʑ]

CUBEO--[”ɺ”, d̠]

DAFLA--[ʌ, i̥] or [ʌ,u̥]

DAGBANI--[”ə”, ŋ͡m]

DAJU--[ɟ, ɾ.]

DAKOTA--[ʃ′, x′]

DANGALEAT--[ɟ, l̠]

DANI--[kʷ, oi]

DIEGUEÑO--[ɮ̻, ɹ]

DINKA--[ɛ̤, ḛ] or [ɛ̤,o̰] or [ɔ̤,ḛ] or [ɔ̤,o̰]

DIYARI--[n., D]

DOGON--[”d”, D]

EJAGHAM--[ʌ, d̠]

EYAK--[”tɬʰ”, ɑ̃ː]

FARSI--[ɢ, a̟]

FE'FE'--[”ɤ ”, ”z”]

FUZHOU--[”tsʰ”, ɶ]

GADSUP--[ß, a̟ː]

GARAWA--[c, t̠]

GBEYA--[ᵑᵐg͡b, n̰]

GEORGIAN--[”ə̆”, ”ɬ ”]

GERMAN--[øː, pf]

GREEK--[ð̻, ɑ]

GUAHIBO--[”ə̃”, r]

GUAJIRO--[”ɺ”, ”r”]

GUAMBIANO--[ß̯, t.s.]

GUARANI--[ŋʷ, ɨ̃]

GUGU-YALANDYI--[”rr”, ɹ.]

HADZA--[!ʰ, ŋ̊!ʔ] or [!ʰ,ŋ̊‖ʔ] or [!ʰ,ᵐpʰ] or [!ʰ,ⁿ̻t̻ʰ] or [‖ʰ,ŋ̊!ʔ] or [‖ʰ,ŋ̊‖ʔ] or [‖ʰ,ᵐpʰ] or [‖ʰ,ⁿ̻t̻ʰ] or [ŋ̊!ʔ,ᵐpʰ] or [ŋ̊!ʔ,ⁿ̻t̻ʰ] or [ŋ̊‖ʔ,ᵐpʰ] or [ŋ̊‖ʔ,ⁿ̻t̻ʰ]

HAIDA--[ŋ, qʷʰ]

HAKKA--[”ⁿd”, ”tsʰ”]

HAUSA--[kʲ′, s̰]

HIGHLAND--[”ð”, n̠̥]

HINDI-URDU--[ẽː, n̻]

HIXKARYANA--[d̠, ɺ.]

HUARI--[ɪ̃, ɺ.]

HUAVE--[”ð”, ᵑgʷ]

HUNGARIAN--[ɟʝ, øː]

HUPA--[hʷ, tʃʷʰ]

IATE--[cʰ, ẽ]

IRAQW--[ɑː, sː]

IRARUTU--[ʝ, ⁿ̻d̻]

ISLAND--[”ɤ̃”, aɯ]

ISOKO--[p˂, v̤]

ITONAMA--[”t′”, ”tʲ”]

JAQARU--[t.s., t.s.′]

JEBERO--[k̰, ɹ]

JIVARO--[”ɾ”, n̲]

K'EKCHI--[b̰, q′]

KAINGANG--[ə̃, ⁿ̠d̠]

KALIAI--[ʍ, ɾ.]

KAM--[ɐ, lʲ]

KANAKURU--[d̰, ᶮɟ]

KANURI--[l., z]

KAREN--[ð, e̠]

KAWAIISU--[”ɾ”, ɣʷ]

KERA--[d̰, dʒ]

KET--[”ɵ”, ”dʲ”] or [”dʲ”,”ɬ ”]

KEWA--[ɸ, ɺ]

KHARIA--[ɦ, ɾ.]

KHASI--[ia, uo]

KIRGHIZ--[ɪ̵, ø]

KLAMATH--[c′, l̥]

KLAO--[ɔ̃, ɟʄ]

KOMI--[ç, ɟʄ]

KONKANI--[ã̟, ɟʑ]

KORYAK--[ß̯, ʄ]

KOTA--[ɾ., s.]

KOYA--[ɐ, ɑ]

KPELLE--[gʷ, k͡p]

KUNIMAIPA--[ɢ, ɾ.]

KURDISH--[əi, lˤ] or [əi,sˤ]

KURUKH--[”õ”, ɟ]

KWAKW'ALA--[n̰, tɬʰ] or [qʷʰ,tɬʰ]

LAME--[ao, eo]

LELEMI--[d., k͡p]

LENAKEL--[ɰ, pʷ]

LITHUANIAN--[æi, rʲ] or [æi,sʲ]

LUE--[kʷʰ, ɣ]

LUGBARA--[d̪̰, ɺ]

LUISENO--[qʷ, s.]

MAASAI--[ɠ, r̥]

MAMBILA--[bv, n̲]

MANCHU--[”ɵ”, ø]

MARI--[ʌ, ʎ]

MAUNG--[ɾ., ɹ]

MAXAKALI--[ⁿdʒ, ɯ̃]

MAZAHUA--[”sʰ”, ɲ̰]

MAZATEC--[ⁿ·d.z., ⁿdz]

MBA-NE--[ᵑ͡ᵐɡ͡b, ⁿdʒ]

MBABARAM--[d̠, ⁿ̠d̠]

MIEN--[cʰ, ŋ̊]

MIXE--[æ, tθ]

MOGHOL--[ʀ, ʄ]

MONGUOR--[ɢ, ʄ]

MORO--[ɺ., ɣ]

MOVIMA--[ɬ, s.]

MUINANE--[rrʲ, ʑ]

MUMUYE--[ŋʷ, ʒ]

MUNDARI--[d̤ʒ̤, n.]

NANAI--[ɪ, ʄ]

NAVAJO--[”dɮ”, ɛ̃ː] or [”dɮ”,õː]

NDUT--[”d”, ᶮɟ]

NEPALI--[d̤., d̤z̤]

NEZ--[n̪̰, qχ]

NGARINJIN--[ɨ, ɹ.]

NIMBORAN--[”ɣ”, ɨ]

NONI--[”d”, ”eː”]

NUNGGUBUYU--[ɰ̟, ɹ.]

NYANGI--[ɠ, s̪]

NYIMANG--[ɟ, ɺ]

OGBIA--[”d”, ”nʷ”] or [”nʷ”,ɐ]

OJIBWA--[ɛ̃ː, ʊ]

PAIWAN--[”dʲ”, ɺ.]

PANARE--[”ə̃”, ”ɣ ”]

PAPAGO--[d., n̲]

PASHTO--[ɑː, ɺ.]

PAYA--[”ɾ”, w̃]

PICURIS--[”ə̃”, ”ɬ”]

PO-AI--[ɛ̆, aɨ] or [ɛ̆,ɪ̆] or [ɛ̆,ɨ̆] or [ɛ̆,ŭ] or [aɨ,ŏ] or [ɪ̆,ŏ] or [ɨ̆,ŏ] or [ŏ,ŭ]

POHNPEIAN--[pʷ, t.s.]

QUECHUA--[”t′”, ʎ]

RESIGARO--[dʲ, n̥]

RUKAI--[t., v]

SALIBA--[”ẽ”, gʷ]

SANGO--[ɺ., ⁿʒ]

SAVOSAVO--[ᶮɟ, z]

SEDANG--[”r̥”, ”r̰”]

SHIRIANA--[”ə̃”, ɯ̃]

SINHALESE--[ɟʝ, ʊ]

SIRIONO--[kʲ, ⁿdʒ]

SOCOTRI--[”ɮ”, ”r̰r”]

SOUTH_KIWAI--[”rr”, ou]

SPANISH--[θ̪, D]

SRE--[ɒ, cʰ]

TAISHAN--[æ, kʷʰ]

TAMA--[ħ, ɟ]

TAMASHEQ--[”eː”, d̪ˤ] or [”eː”,t̪ˤ] or [”eː”,z̪ˤ]

TAMPULMA--[ŋ͡m, ɟ]

THAI--[ɾ, t̪s̪ʰ]

TIGAK--[ɐ, ɮ]

TIGRE--[a̟ː, ts′]

TIWI--[$^{\text{n}.\mathit{66}}$t., ɹ.]

TOL--[t̰s̰, tsʰ]

TRUMAI--[m̥, xː]

TSESHAHT--[”tɬ”, ʔˤ]

TSIMSHIAN--[ɰ̰, kʲ′]

TSOU--[”l̰”, ʉ]

TULU--[s., ʊ]

TURKISH--[ɰ̟, ʏ]

TUVA--[β, ɤ]

TZELTAL--[t̪s̪′, y]

UPPER--[”ə”, ”tɬ′”]

WARAO--[β̞, ɺ]

WARIS--[”ⁿd”, ɒ]

WINTU--[q′, ɹ]

WIYOT--[ɹ., ɹ]

YAGUA--[ʊ̃, ɨ̃]

YAWA--[nʲ, ɾ]

YAY--[ua, ɯa]

YESSAN-MAYO--[ᵑgʷ, ɒ]

YOLNGU--[ɐ, ɾ.] or [ɹ.,ɾ.]

YUCATEC--[”t′”, ɛ̰] or [”t′”,ɔ̰] or [”t′”,ṵ]

YUCUNA--[ɺ, tʰ]

YULU--[ɗ̪, ⁿ̪z̪]

YUPIK--[”ɹ̊”, ɣʷ]

3-segment niches

ABIPON--[ʁ, ɨ, q]

ALABAMA--[”ɬ”, ɸ, m]

AMO--[ɐ, k͡p, ts]

ANDAMANESE--[”d”, ”ɹ̊”, æ]

ARABELA--[”s”, ʃ, au]

ASMAT--[”ə”, ɔ, ʝ]

AUCA--[”ẽ”, æ̃, ı̃]

BAKAIRI--[”d”, ʒ, ẽ]

BAMBARA--[ɔ̃, ẽ, D]

BARASANO--[”ẽ”, ɨ̃, ɾ]

BELLA--[”tɬ′”, c′, ts′]

CACUA--[ə̃, ɛ̃, ʍ]

CARIB--[ɐ, β, ɯ]

CHEROKEE--[”ə̃”, d, dz]

DADIBI--[”ẽ”, tʰ, ũ]

DAGUR--[”l”, dʒ, y]

DIOLA--[ə, ɟ, ɾ]

DIZI--[”ts′”, ɐ, β]

DOAYO--[”d”, ɔ̃, k͡p]

FINNISH--[æ, lð, ø]

JAPRERIA--[”t”, ɨ̃, ɺ]

KADUGLI--[”e”, ɗ̪, ʄ]

KALA--[”ə”, n̲, z]

KOMA--[ɓ, p′, s′]

KPAN--[ɔ̃, dz, ⁿd]

KULLO--[”d”, ”d <”, ”ts′”]

KWAIO--[ŋʷ, ᵑgʷ, xʷ]

LAHU--[ə, ɨ, qʰ]

LUA--[”ə̃”, ”ẽ”, ɗ]

LUSHOOTSEED--[ʊ, dz, tɬ′]

MABA--[d., ɟ, z]

MAIDU--[c′, cʰ, ɗ]

MARGI--[ɗ, dz, ɮ]

MIXTEC--[”õ”, ð̪, kʷ]

MURINHPATHA--[d, ɾ., t.]

MURSI--[”d”, θ̪, ɟ]

NGIYAMBAA--[ɹ., D, t̠]

NICOBARESE--[ə, ɒ, ɯ]

ORMURI--[”dz”, ß̪, a̟]

POMO--[ʊ, p′, x]

SENECA--[”dz”, ”o”, ɛ̃]

TABI--[θ̪, ð̪, ɟ]

TAROK--[”d”, ʒ, k͡p]

TEMNE--[ɑ, g͡b, t̪]

TICUNA--[”õ”, ”ɾ”, ɨ̃]

TIDDIM--[l̰, w̰, z]

TIRURAY--[ŋ, ɸ, ɨ]

TLAPANEC--[dʒ, ĩ, ts]

TONKAWA--[”ts”, eː, xʷ]

TOTONAC--[aː, ɬ, uː]

USAN--[ʌ, d, ⁿd]

WAPPO--[”l̰”, ”ts′”, m̰]

WEST--[”r”, ɸ, n̲]

WIK-MUNKAN--[ɔ, ʊ, t̠]

YANA--[”rr”, tð′, ui]

YAREBA--[”e”, ɸ, dz]

YORUBA--[dʒ, k͡p, ɾ]

4-segment niches

ACHUMAWI--[”ə”, ”rr”, χ, dʒ]

BURARRA--[ɛ, n., ɹ., u]

CHATINO--[”d”, ”ẽ”, ʃ, ĩ]

CUNA--[”d”, ”ɾ”, gʷ, tʃ]

FASU--[”ẽ”, ɸ, ɾ, w]

FUR--[”ə”, aː, d, r]

JINGPHO--[”dz”, ”rr”, ”tʰ”, ”ts”]

JOMANG--[ʊ, d, d̪, ɟ]

KALKATUNGU--[lð, ʎ, ɹ., ɾ]

KEFA--[”rr”, ”t”, f, p′]

KUNAMA--[”o”, ʊ, ɲ, r]

QAWASQAR--[”ə”, t′, tʃ′, x]

QUILEUTE--[l, qʷ′, tɬ, ts′]

SENTANI--[”ə”, ”d”, a̟, f]

XIAMEN--[”dz”, ”t”, ”tsʰ”, ɔ]

ZUNI--[”ɬ”, ”t”, ”tsʰ”, kʷ′]

5-segment niches

BODO--[”ə”, ŋ, d, ɾ, z]

BORORO--[”ə”, ”d”, ”rr”, dʒ, ɨ]

GWARI--[”e”, ”o”, ĩ, k͡p, z]

MARANAO--[”o”, ŋ, h, ɨ, ɾ]

MOR--[”e”, ”rr”, ʔ, ß, w]

SA'BAN--[”ə”, ŋ, d, ɨ, ɾ]

SHASTA--[”rr”, ”t”, p′, tʃ′, x]

TUNICA--[e, r, tʃʰ, tʰ, w]

6-segment niches

BARIBA--[e, k͡p, ɾ, t, ũ, z]

IVATAN--[”ə”, d, h, r, t̪, v]

SIERRA--[ʔ, ɛ, ŋ, ʃ, ɨ, s̪]

TAGALOG--[ɪ, ŋ, n̪̤, ɾ, s, t̪]

TEMEIN--[ɟ, n̪̤, ɲ, r, s, w]

7-segment niches

TETUN--[”d”, ”e”, ”r”, ”s”, ʔ, f, p]

II. Languages discriminated by negative niches

3-segment niches

BIROM--[c, k͡p, not-ɲ]

EFIK--[k͡p, not-g, not-”d”]

NASIOI--[”ɾ”, not-w, not-”l”]

RORO--[ɾ̪, ”o”, not-”n”]

4-segment niches

BANDJALANG--[n̲, r, not-k, not-”d”]

CAMPA--[t̠, e, ß, not-”e”]

KOIARI--[ð̪, h, ɾ, not-”l”]

PIRAHA--[not-^{m}b, not-j, not-m, not-u]

ROTOKAS--[D, t, ß, not-ʔ]

SUENA--[”dz”, ”ɾ”, not-”ɾ”,not-”e”]

WESTERN DESERT--[ɾ,̠l, not-b^{m}, not-ʔ]

5-segment niches

CHUAVE--[f, ɾ, not-p, not-h, not-ŋ]

DERA--[”ə”, ŋ, ”d”, not-”l”, not-”d^{j}”]

IWAM--[”ə”, ŋ, not-b, not-l, not-”n”]

MOXO--[ɛ, n̲, ɾ, ß, not-”o”]

NUBIAN--[”l”, ”e”, d̪, ɲ, not-”d”]

6-segment niches

BATAK--[”l”, ”s”, ”o”, ɛ, dʒ, not-”e”]

BISA--[v, z, not-x, not-h, not-tʃ, not-ɛ]

IBAN--[”ə”, ”e”, dʒ, r, not-”s”, not-ʃ]

NERA--[”d”, not-p, not-ʔ, not-x, not-n̲, not-”t′”]

SONGHAI--[”e”, o, not-ɛ, not-h, not-ʃ, not-”o”]

7-segment niches

AINU--[ɾ, not-l, not-b, not-”e”, not-ʔ, not-ʃ, not-e]

HAWAIIAN--[”l”, ”n”, ”o”, ɛ, ʔ, h, not-”e”]

TAORIPI--[”e”, ”l”, ”s”, not-b, not-w, not-x, not-ß]

8-segment niches

YAQUI--[”e”, ɾ, ʔ, u, not-ã, not-ʃ, not-ɬ, not-”d”]

III. Two indistinguishable languages

Dyirbal and Yidiny cannot be distinguished from one another, but the shared niche [ɟ,ɹ.] distinguishes them from all remaining languages.

Notes

**Acknowledgments*. Raúl E. Valdés-Pérez was partly supported by grant #IIS-9819340 from the USA National Science Foundation (NSF) and Vladimir Pericliev by a supplement to the latter grant from the NSF Division of International Programs and grant #I-813 from the Bulgarian Ministry of Education and Science.

[1] The languages using negative niches mentioned above require from 3 to 8 segments for discrimination, cf. Appendix.

[2] The demonstrative (non-probabilistic) validity of this statement is obvious in the limiting case of the smallest, 1-segment, niches, which will, of necessity, comprise only the most infrequent segments, viz. the segments that occur only once and in the languages they actually discriminate. A more common segment cannot do the job. We nevertheless express the statement more tentatively as a tendency since in larger niches things are not that clear. The reason is that one can imagine a situation in which a segment set of some size includes uncommon segments but still cannot act as a niche, while a set of the same size includes more common segments but is discriminative for a language. Extreme situations like this imaginary one aside however, it could be expected that niches are more likely inhabited by segments of lower frequencies.

[3] The fact that 2- and 3-segment niches might contain segments of frequency=1, of course, implies that these (non-anomalous) segments in the respective languages are uncontrastable with other segments from other languages; hence they are insufficient---by themselves---to achieve 1-segment discrimination with all remaining languages.

[4] Cf. Maddieson (1984: 14-16) for a discussion of some universals of this logical form, which are interpreted as prohibitions on the co-occurrence of phonetically similar segments in the inventory of one language.

References

Croft, W. (1990). *Typology and Universals*. Cambridge: Cambridge University Press

Dryer, M. (1992). The Greenbergian word order correlations. *Language* 68, pp. 81-138

Ferguson, Ch. (1978). Historical background of universals research. In: J. H. Greenberg (ed.), *Universals of Human Language, Method and Theory*, vol. 1, pp. 7-32. Stanford: Stanford University Press

Greenberg, J. H. (1966). Some universals of grammar with particular reference to the order of meaningful elements. In: J. H. Greenberg (ed.), Universals of Language, pp. 73-113. Cambridge, Mass: MIT Press

Greenberg, J. H. (1973). The typological method. In: Th. Sebeok (ed.), *Current Trends in Linguistics. Diachronic, Areal and Typological Linguistics*, vol. 11, pp. 149-194. The Hague: Mouton

Hawkins, J. (1983). *Word Order Universals*. New York: Academic Press

Hombert, J.-P. &. Maddieson, I. (1999). Rare segments and automatic language identification. *Eurospeech*, Budapest

Laver, J. (1991). *Principles of Phonetics*. Cambridge: Cambridge University Press

Lindblom, B. & Maddieson, I. (1988). Phonetic universals in consonant systems. In: L. M. Hyman & C. N. Li (eds.), *Language, Speech and Mind: Studies in Honor of Victoria A. Fromkin*, pp. 62-80. New York: Routledge

Maddieson, I. (1984). *Patterns of Sounds*. Cambridge: Cambridge University Press

Maddieson, I. & Precoda, K. (1991). Updating UPSID. *UCLA Working Papers in Phonetics* 74, pp. 104-114

Murdock, G. P (1970). Kin term patterns and their distribution. Ethnology 9, pp. 165-207

Pericliev, V. & Valdés-Pérez, R. (1998a). A procedure for multi-class discrimination and some linguistic applications. In: *Coling-ACL'98, 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, vol. II, pp. 1034-1040, August 10-14, Université de Montréal, Montreal, Quebec, Canada

Pericliev, V. & Valdés-Pérez, R. (1998b). Automatic componential analysis of kinship semantics with a proposed structural solution to the problem of multiple models. *Anthropological Linguistics* 40, pp. 272-317

Pericliev, V. & Valdés-Pérez, R. (1998c). A discovery system for componential analysis of kinship terminologies. In B. Caron (ed.), *Actes du 16è Congrès International des Linguistes* (Paris, 20-25 juillet 1997). CD-ROM published by Pergamon/Elsevier

Pericliev, V. (1999). The prospects for machine discovery in linguistics. *Foundations of Science* 4, pp. 463-482

Valdés-Pérez, R. & Pericliev, V. (1997). Maximally parsimonious discrimination: a task from linguistic discovery. In: *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, pp. 515-520. Menlo Park, Calif.: AAAI Press

Valdés-Pérez, R. & Pericliev, V. (1999). Computer enumeration of significant kinship universals. *Cross-Cultural Research* 33, pp. 162-174

Valdés-Pérez, R., Pereira, F. & Pericliev, V. (2000). Concise, intelligible, and approximate profiling of multiple classes. *International Journal of Human-Computer Studies (Special Issue on Machine Discovery)* 53, pp. 411-436

Table 1. Some statistics on positive niches

Size of niches	Number of languages distinguished by a niche of some size[a]	Average sizes of reduced segment inventories of languages distinguished by a niche of some size	Average number of alternatives within niches of some size
1	124	37.1	2.8
2	213	29.5	19
3	57	25.4	29.9
4	16	24	30.1
5	8	20.6	13.1
6	5	23	18
7	1	18	4

[a] Two languages are indistinguishable from one another, and 25 other languages can only be differentiated with negative niches.

Table 2. Distribution of vowels and consonants computed from positive niches in the Appendix

Size of niches	Structure of niches	Number of exemplifying niches
1-segment niches	C	240
	V	102
2-segment niches	C + C	132
	V + C	82
	V + V	33
3-segment niches	V + C + C	25
	C + C + C	17
	V + V + C	13
	V + V + V	2
4-segment niches	V + C + C + C	5
	V + V + C + C	5
	C + C + C + C	6
5-segment niches	C + C + C + C + C	1
	V + C + C + C + C	3
	V + V + C + C + C	3
	V + V + V + C + C	1
6-segment niches	C + C + C + C + C + C	1
	V + C + C + C + C + C	2
	V + V + C + C + C + C	2
7-segment niches	V + C + C + C + C + C + C	1
	Total	676