Syncretism Distribution Modeling: Accidental Homophony as a Random Event*

Uli Sauerland and Jonathan David Bobaljik

Harvard University/ZAS and University of Connecticut

Abstract

The morphological analysis of paradigms generally proposes a distinction between accidental and systematic homophony. No specific assumptions are usually made about the distribution of accidental homophony, though. Therefore current assumptions cannot prove satisfactorily what should be regarded as systematic in morphology. We propose that accidental homophony should be assumed to be a random event in the statistical sense with a constant probability across languages and across paradigms. This approach allows us to assign a likelihood to any actual typological distribution of syncretism given a morphological analysis. And by computing such likelihoods for a range of analyses, we can then apply maximum likelihood analysis to determine the best analyses. Hence, the statistical foundation allows us to empirically test morphological analyses that include accidental syncretism. In this paper, we primarily introduce the conceptual and mathematical foundations of a statistical modeling technique, Syncretism Distribution Modeling, and show how it overcomes the problem of accidental homophony. In addition, we apply the technique to show that person paradigms must involve both accidental homophony and systematic syncretism.

*For helpful comments, we are grateful to Daniel Harbour, Irene Heim, Roni Katzir, Gereon Müller, Kazuko Yatsushiro, and to audiences at GLOW 35 at Potsdam University, GLOW Asia 9 at Mie University, MIT, and the Universities of Connecticut, Iceland, Jerusalem, Leipzig, and Tohoku (Sendai). The research reported on in this work is in part supported by funding from the Humboldt foundation, the German Ministry for Education and Research (BMBF), grant number 01UG0711, and the US National Science Foundation, grant #BCS-0616339.

In print in: Nobu Goto, Koichi Otaki, Atsushi Sato, and Kensuke Takita, Proceedings of GLOW in Asia IX 2012: XX–XX

1 Introduction

1.1 The Problem of Accidental Homophony

To draw the distinction between accidental homophony and systematic syncretism poses a general problem to linguists of all stripes. There are some clear cases of both types. For example, the English words *bank* 'financial institution' and *bank* 'side of a river' are a textbook example of *homophony*: two distinct words (or morphemes) that, due to the vagaries of history, happen to share the same sound. The different meanings are readily seen in translation, where other languages have two different words for these concepts, e.g., in German *Bank* ('financial inst.') and *Ufer* ('shore'). Consider now the English second person pronoun *you*. This single word in English also corresponds to multiple words in other languages, for example, German uses *du* for a singular addressee, and *ihr* for plural addressees. One possible analysis of English would be to posit two words *you* that share the same sound. Yet rather than positing homophony, it is common to think of English *you* as a case of under-differentiation: English pronominal morphology neutralizes a contrast that is made in other languages, namely, that between singular and plural second persons. The term *syncretism* refers, in the morphological literature, to such under-differentiation in a paradigm, where a single *exponent* (i.e., phonological unit) corresponds systematically to multiple cells. In general, though, there is no theory-independent procedure to determine whether a given case of homophony is due to accidental homophony or systematic syncretism (see also Harbour 2008).

In this paper, we address the distribution of homophony patterns in person paradigms. The English and German personal pronoun paradigms are shown in figure 1. Both languages have three grammatical persons and two numbers, yielding a six-celled *paradigm*. However, whereas English has five distinct (combinations of) sounds for the six cells, German has six. The relation of homophony between cells defines for each language a partition of the cells into equivalence classes. In the following, we use the term *D-Partition* for the partition of the cells of a paradigm defined by homophony. The d-partitions of English and German personal pronouns are indicated by the abstract diagrams below the paradigms in 1. In the following, we are only concerned with the d-partitions, never with the actual phonological content of the cells. Thus, Itelmen, an indigenous language of the Kamchatka peninsula in Russia, shows a 6-way contrast like German. Although the actual phonological forms are entirely unrelated to their German counterparts, we consider Itelmen and German to constitute distinct tokens exemplifying the same d-partition. We note, in addition, that in thinking of paradigm distribution as a partition problem, we ignore the geometric layout of the standard presentation of a paradigm. A paradigm space is an unordered set of feature combinations, and the partitions are simply the possible subsets of this set. There is no meaningful sense in which cells are adjacent to one another or not. We use a constant layout only to facilitate easy identification of correspondences among different partitions.

Our interest in d-partitions in person paradigms is due to the observation that many possible d-partitions are very rare or don't occur at all, while others are very frequent crosslinguistically (Forchheimer 1953, Cysouw 2003). The number of possible partitions

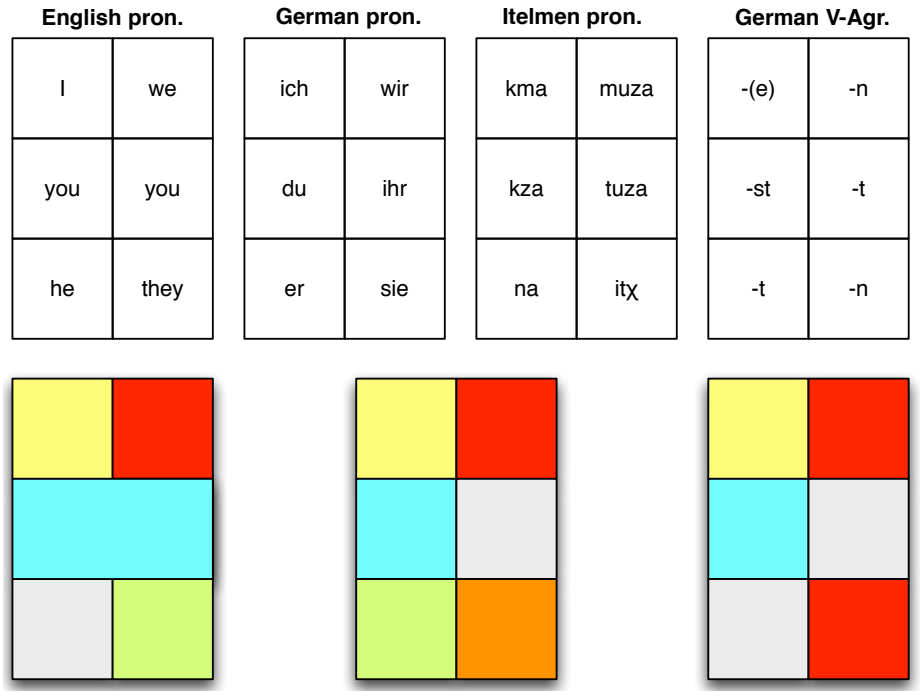

Figure 1: Paradigms and D-Partitions for Person in English, German and Itelmen

of an n-membered set is the Bell number B_n, which grows exponentially with n. For four cells, there are $B_4 = 15$ different d-paradigms, any six-celled paradigm has $B_n = 203$ possible d-partitions, and for eight cells there are 4140 possible d-partitions. We discuss in more detail in the following sections evidence that indicates that the frequency distribution of d-partitions across languages requires an explanation. There have already been some attempts to explain this extremely skewed distribution, though these have been for the most part qualitative, rather than quantitative (but see Pertsova 2011). Accounts in the morphological literature posit a universal feature inventory, which allows observed (and in particular common) syncretic patterns (d-partitions) to be described as neutralizations of underlying contrasts (see Bobaljik 2008a, Harley and Ritter 2002, Wechsler 2010). Such approaches therefore argue for language-universal constraints against some types of syncretism. But in attempting such a study, a significant problem presents itself, namely the problem of distinguishing the accidental from the systematic–homophony from syncretism.

As we mentioned above, it is not clear how to draw the distinction between homophony and syncretism in general. Consider one relevant case where different proposals have been made: the German verbal agreement paradigm, also shown in figure 1. Unlike in the German pronouns, in the verb endings there are only 4 phonologically distinct endings spread over the six cells; the first and third person plural endings are identical (*-en*), as are the second person plural and third person singular (*-t*). In principal, both of these two cases of homophony could be either accidental or syncretism. For example, it may be that the

German speaker's mental grammar includes six exponents for the present tense endings, but that it just so happens that some of the listed endings have the same phonological realization: (1sg = *-e*, 1pl = *-en*, ... 3pl = *-en*, etc.). On this view, the underlying structure of the paradigm doesn't correspond to the d-partition, but is the same as that of German and Itelmen pronouns. In what follows, we will use the term *M-Partition*, to refer to the underlying, abstract grammatical structure of the paradigm. Only if there is no accidental homophony, the m-partition is the same as the d-partition. In the analysis just discussed, the m-partition is the same as the paradigm space, that is, the maximal partition of that space. A second conceivable m-partition for German verbal agreement is the d-partition. In addition, two intermediate m-partition exist in this case. All four possible m-partitions in this case are shown in figure 2.[1] We compare partitions in terms of their *coarseness*. Partition m_1 is at least as coarse as m_2 if any two members of the same equivalence class of m_2 also belong to the same equivalence class in m_1. Coarseness partially orders the m-partitions in figure 2, with the coarsest m-partition on the right being coarser than the other three, and the finest on the left finer than the other three. But the middle two m-partitions in figure 2 are not ordered relative to each other by coarseness.

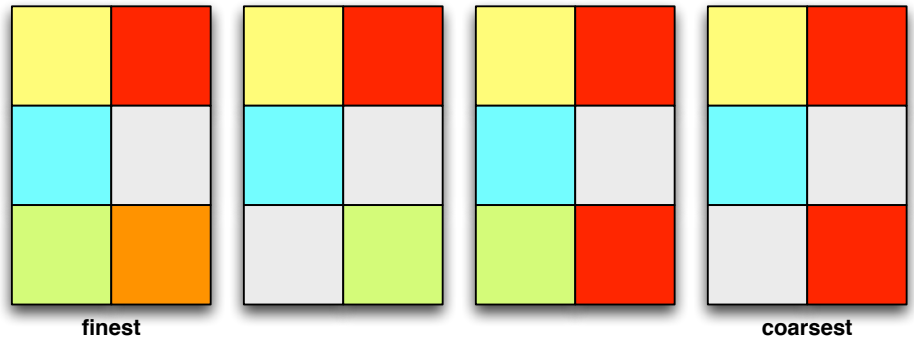

Figure 2: Possible m-partitions for German Verb Agreement

The decision of which m-partition is correct in a case like German is usually made on the basis of theory internal criteria. As we mentioned, at least two morphological analyses for German verb agreement have been offered in the literature corresponding to the two m-partitions on the right in figure 2. The two analyses of the German facts, cast as ordered rules of exponence (see Anderson 1992, Stump 2001), are given in (1a) and (1b). Analysis (1a) is an instance of the third m-partition of figure 2, while (1b) instantiates the coarsest m-partition. The two analysis also use a different feature logic. Analysis (1a) only uses positively specified features, while analysis (1b) allows appeal to both negatively and positively specified features to admit [-2] as a feature that includes 1st and 3rd person. For this reason, the analysis can be successful without appeal to accidental homophony.

[1] In general, the number of possible m-partitions for a given d-partition is the product $\prod_e B_{c(e)}$ where e ranges over the exponents of the d-partition and $c(e)$ is the number of cells where exponent e occurs.

(1) a. 2 pl → -t
pl → -en
2 → -st
3 → -t
default → Ø

b. [-2, +pl] → -en
[+2, -pl] → -st
[+1, -pl] → Ø
[-1] → -t

We have kept above to a simplified version of the general problem. As paradigm spaces (and the features involved) expand, the problem of distinguishing the accidental from the systematic grows quickly. Homophony happens. It is thus impossible to know a priori what the correct m-partition is for any given d-partition. In the simple case of the German present tense that we have just looked at, the facts are clear, but we have given three analyses invoking 4, 5, or 6-membered m-partitions. This case also makes clear that the spectre of circularity looms large here, as we seem to be trading a more complex analysis for a reduction of accidental homophony.

There is no agreement in the field as to which is the correct m-partition and how to decide such a question. As Harbour (2008) points out, linguistics has not yet come to grips with the problem of accidental homophony. In analyzing identity of form within paradigms, some authors even deny that there is any internal feature decomposition at all, and thus effectively treat all formal identity as a form of homophony (Cysouw 2003). In some sense, the other extreme position is advocated by Halle and Marantz (2008) who seek to *avoid accidental homophony and maximize generalization*. Concretely they propose to count across languages the number of cases of accidental homophony a morphological analysis requires minimally, and always choose the analysis with fewer instances of accidental homophony. This proposal however is problematic for at least three reasons: For one, the analyses compared may differ substantially in their internal complexity and it would not seem reasonable to adopt a much more complex analysis (for example, one with negative feature values in addition to positive ones) only to avoid a single instance of accidental homophony. Secondly, the choice of analysis could depend very strongly on the languages for which data is available. The oversight of a single language that would require accidental homophony may lead us to prefer a wrong analysis. Finally, there might be cases where one analysis that requires least homophony requires always homophony between the same two cells. A different analysis, however, leads to more instances of homophony overall, but these are distributed evenly over different pairs of cells. In such a case, we would intuitively prefer the latter analysis, but Halle and Marantz (2008) would have us advocate the former. The problem of accidental homophony also cannot be addressed in the same way as some other problems in morphology. One difficulty is that paradigms are typically not just finite, but relatively small and closed. To test the productivity of other morphological processes, researchers have relied on variants of the *wug*-test of Berko (1958): productive processes are extended by speakers to novel or nonce forms they have never heard before, and thus could not have learned. The *wug*-test works fine to demonstrate that past tense is productive, but doesn't help with our problem: When it comes to person, there is no reasonable way to simply add a new feature or grammatical person to the system.

In the rest of this paper, we propose a new method to evaluate morphological models. The underlying intuition is that accidental homophony can be viewed as a random event. This gives us a lot of predictions, because a random event is a statistical concept: A random event is very likely an even distribution of accidental homophony and very unlikely to have an uneven distribution. For a morphological model and a fixed set of observations, we can then distinguish between the homophonies that can be analyzed as systemic syncretism in the morphological model and the homophonies that must be accidental. Then we can test whether the distribution of the accidental syncretism is likely or unlikely given our assumption that it is a random event. By this method, we can for different models of morphology compute how likely they predict a set of actual observations to be. Since we assume that the real world that the observations are taken from is fully explained by the model and the random variable of accidental homophony, we would then prefer models that result in a comparatively high likelihood for the set of observations we are looking at. In this way, we can select from a number of morphological models the one that assigns the highest likelihood to the real world. In a very loose sense, what we propose is akin to the following exercise: imagine a situation in which a black box contains an unknown number of regular 6-sided dice, that can be thrown, with the total of each throw reported. What we observe directly is some distribution of totals per throw: throw one, total 8; throw two, total 6, throw three, total 7, throw four, total 7. While this observation is consistent with the model where there are six dice in the box, the model with just two dice is clearly is the one that best fits the observations. The problem of finding a good morphological model is more complex, of course, but having the dice-detecting analogy in mind may help to think about what we are doing below.

1.2 The Cysouw Data-Set

Our analysis in this paper makes use of the substantial data set available for the domain of person paradigms from a large, cross-linguistic sample originally collected and analyzed by Cysouw (2001, 2003). We now first introduce this data set, before returning to the issue of accidental homophony and the theoretical tools we will make use of. As far as we know Cysouw's data-set is the largest collection available, and although there are limitations, it has several good properties. Cysouw's data-set is primarily gathered from existing grammatical descriptions, established by many others. However, the coding of person is well-described, with broad consensus on the observed categories. So, in our view, Cysouw's data are of high reliability, though that is most likely also the case for other datasets in this area. A more significant advantage is that Cysouw, as already mentioned, arrives at conclusions that are quite different from the ones we will argue for: namely, Cysouw argues that homophony is not indicative of an underlying feature structure (in common understanding, it is synchronically accidental), while we will show here that such an analysis is inconsistent with Cysouw's data.[2] So, we can be sure that the dataset is not selected with a bias

[2]Note that Cysouw (2011) can be read as retreating somewhat from his earlier view, acknowledging to some degree the independence of number from person.

towards our conclusion. For these two reasons and also mere convenience, we chose to use Cysouw's data-set.

Cysouw's data set consists of a survey of some 265 tokens of paradigms (i.e., d-partitions in our terms) of person marking from the world's languages, representing 61 types, and indicating the relative frequency of each type in the sample.[3] The dataset includes both independent pronouns and inflectional elements (bound affixes). It is by far the most comprehensive survey of its sort, however words of caution are in order regarding interpretation of the quantitative data. Of particular relevance is a skew in favour of the over-representation of rare paradigms. Thus: "[e]very single example of a rare paradigmatic structure has been included in the sample to show the inherent variability of human language ... On the other hand, not every case of the commonly occurring paradigmatic structures is included. After a representative group of a specific common paradigmatic structure has been described, [Cysouw] stopped the collection of more exemplars of such common patterns." (Cysouw 2003, 22-23). In addition, to control (in part) for the likelihood that shared paradigm structures among related languages arise from common ancestry, rather than from general properties of the person system, Cysouw limited the counts of any individual d-paritition to no more than 4 from a given family (Cysouw 2001, 341). Acknowledging that this move limits our confidence in the results we may obtain, we press forward with this data set in any event, as a more accurate sample will have only a more pronounced divide between the common and the rare, and should thus produce cleaner results. Other concerns about the data set concern independence of the data points and are identified below.

Cysouw codes, descriptively, for an 8-cell parameter space of person and number, rather than the six-cell space familiar to speakers of Western European languages and presented with reference to German above. The underlying person categories are speaker/author [1], hearer/addressee [2], the combination of the two [1+2] ("first person inclusive"), and non-participant [3], and these may be singular (minimal) or non-singular (a group containing the designated referent. A language with the full 8-way contrast is Ilocano as show in figure 3, where also English and German as eight-cell paradigms are shown for comparison.

The minimal inclusive cell has been the subject of much discussion since it was first identified (Thomas 1955). Minimal inclusive pronouns refer to one speaker and one hearer (you and me) and no others, and form a number pair with a corresponding plural (a group containing this duo). These pronouns are not grammatically duals (see Cysouw 2003, Bobaljik 2008a for reviews of the literature).

This 8-cell paradigm space marks the maximal set of person contrasts in the world's languages. There is never a person-contrast of more than 4 persons in any one grammatical number, and moreover, if there is a four-way contrast, it is always the one in figure 3. Additional contrasts involve cross-classification with independent features such as gender,

[3]We rely primarily on Cysouw (2001), since quantitative information is provided in the appendix to that work. Cysouw (2003) offers a few minor corrections and reclassifications of some of the data. At this preliminary stage, our main purpose is to test the model over an existing data set; in ongoing work, we are collecting a more careful sample, allowing for more robust results than Cysouw's.

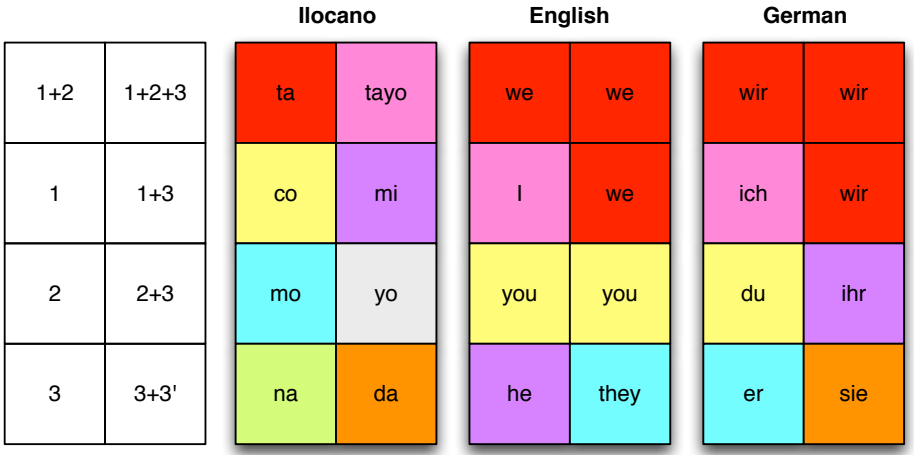

Figure 3: Full Eight-Cell Person Paradigms of Ilocano, English, and German

case, etc.[4] All known paradigms (d-partitions) can be derived from this as neutralizations of one or more contrasts. For example, English and German lack an inclusive-exclusive contrast, thus three of the 8 cells have an identical form (e.g., English *we*).

We note off the bat that Cysouw's data set, though constructed with an eye to maximizing the representation of rare types, includes only 61 d-partition types out of 4 140 ($= B_8$) logically possible types.

Cysouw also presents separately the results for a smaller space, the top four cells of the eight in figure 3, which correspond to the various first-person markers (inclusive vs. exclusive $\times$ singular/minimal vs. plural). This is a smaller set of possibilities: there are only 15 partitions of a 4-membered set–restricting investigation to this smaller paradigm space will yield more tractable computations in what follows. We show this data in figure 4. Cysouw presents specific numbers for the occurrences of each of the rare patterns, but, as noted above, capped his collection of examples of the common patterns and does not provide specific counts. We have used 15 as a proxy where Cysouw speaks of "dozens", consistent with Cysouw's methodology of low-balling the counts of the exceedingly common patterns and artificially amplifying the relative frequency of rare paradigm types. With this restriction, we retain the data from 83 paradigms. The distribution is still an unusual distribution: 5 d-partitions have 15 occurrences each, while another 5 d-partitions don't occur at all, as Cysouw already notes.

[4]This is by know means a logical necessity; there have been proposals for additional contrasts, famously between multiple hearers [2pl] versus hearer and others [2+3,pl]; yet no example of such putative contrasts has survived scrutiny (Cysouw 2003, Simon 2005, Bobaljik 2008a).

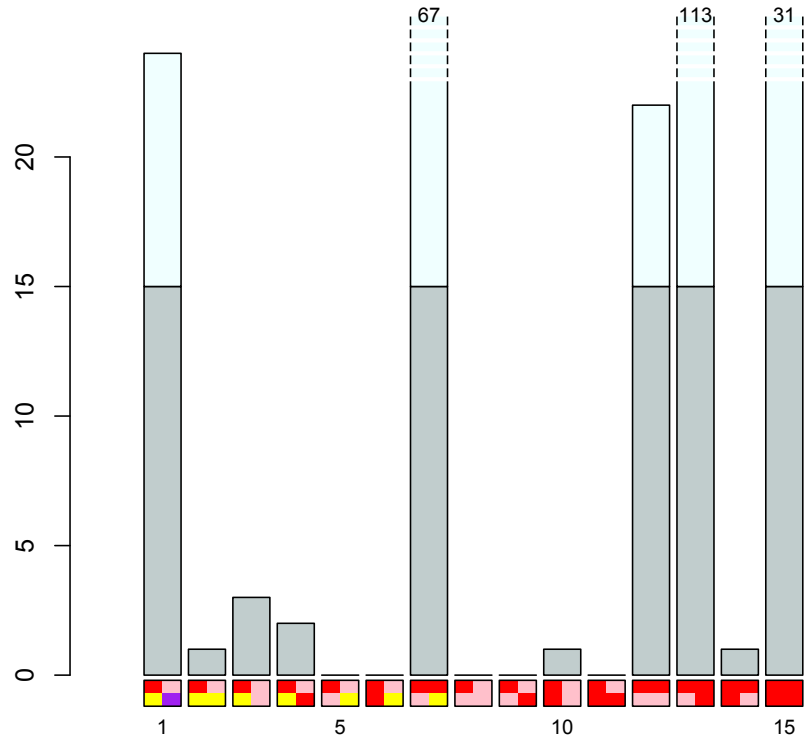

Figure 4: frequencies of First Person d-partitions in Cysouw's dataset

2 Accidental Homophony as a Random Event

In this section, we introduce our main assumption: that accidental morphology is a random event in the statistical sense. Furthermore we show how our main assumption can be used in morphological analysis, specifically by assuming that the probability of accidental homophony is constant across a paradigm and across languages.

2.1 The Rate of Accidental Homophony

Now, consider again accidental homophony. Our proposal is that accidental homophony should be treated as a random event in the comparative typology of paradigms. The major motivation for this proposal is that it captures the intuitions traditionally connected with the notion of accidental homophony; namely, the intuition that through an accidental development in language history—for example, a sound change or the introduction of a loan word—two distinct morphemes end up sounding the same, even though the grammar doesn't allow a meaningful identification of the two morphemes. The factors bringing about accidental homophony are therefore regarded as historical accidents independent of the paradigm structures examined. The assumption that accidental homophony is random doesn't entail that it could not be caused. Also random noise in scientific experiments is

usually taken to have some cause: In scientific experiments, we assume that besides some variables that are controlled for a wealth of known and unknown additional factors affect the result. Therefore, the measurement may differ even if the variables that are controlled for are constant. But as long as it is justified to assume that the other causes are randomly interfering with the ones we control for, we can apply statistical computations to say something about the effect of the controlled variables. We propose that, in the same way, our determination of paradigm structure is affected randomly by accidental homophony.

Our proposal that accidental homophony is a random event makes it possible to approach accidental homophony with statistical methods. Statistical methods rely on the insight that, while a single observation of a random event is not informative, repeated observation usually is: the cumulated outcomes generally show a highly regular behavior. For this behavior to emerge, the different observations must be independent of each other in the sense of probability theory. In essence this means that all the prior observations must not affect the next one like a dice roll doesn't depend on previous rolls.

To apply statistical concepts to accidental homophony, we assume that the basic event is the following choice: Do two distinct morphemes A and B accidentally sound the same or not. We assume that accidental homophony in this case has probability h, the overall *rate of accidental homophony.* If the value of h was 0.1, the two randomly chosen morphemes (in a paradigm) A and B would have a 1 in 10 chance of sound the same and a 9 in 10 chance of sounding different. Only if h was 0, would accidental homophony be completely impossible.

2.2 Two Independence Assumptions

We make two further assumptions as working hypotheses: (i) independence across the paradigm, and (ii) independence across languages. These assumptions mean that for any two morphemes of a particular language from a specific paradigm, h is the likelihood of the two morphemes sounding the same. For the example of person morphology, consider first languages that have an m-partition with four first person morphemes 1, 1+2, 1+3, and 1+2+3. Then h is the chance of 1+2+3 sounding the same as 1 as well as that of 1+2 and 1 sounding the same, and so on for any other selection of two morphemes from the m-partition. Compare this now with languages that have an m-partition with only two morphemes, 1 and [1+2, 1+3, 1+2+3]. Then the same value h is the likelihood of these two morphemes sounding the same.

Both independence assumptions may turn out to be false after investigation, but they are the right points of departure. In the absence of evidence showing an quantifiable effect of other factors, the independence assumptions should be our null assumptions. If at some point in time the effect of our other factors on h can be quantified, the techniques we develop here could be extended to accommodate this in the model. For now, we think both assumptions can be maintained, though they are to some extent idealizations. Consider first independence across languages. Though languages as a whole are certainly not independent of one another, there is to our knowledge no evidence that the distribution of accidental homophony is similar across distinct, related languages. Still care must be taken

in the selection of a language sample to ensure that independence of accidental homophony across languages is a defensible assumption, but the Cysouw dataset that we use satisfies these criteria. A second way in which independence across languages is an idealization is that the likelihood of accidental homophony may depend on the phonological inventory of a particular language. Intuitively it seems plausible that a language with a small inventory of possible syllables should have a greater chance of accidental homophony than a language with a large inventory. However, we are not presently aware of any evidence actually showing such an influence, and therefore feel that it would not be justified to include phonological factors in our statistical model at this point. For independence across the paradigm, it is possible that for example, morpheme frequency affects the likelihood of homophony (William Snyder, p.c.). But, in this case too it isn't sufficient to only look at d-partitions and it therefore is difficult to actually be sure of the plausible effect. So at this point, we resist the inclusion of additional free parameters unless their influence has actually been empirically demonstrated.

3 Accidental Homophony and Morphological Models

3.1 Morphological Parameter Spaces

In the following pages, we present a statistical technique to develop the independence assumptions into a tool to evaluate morphological theories against typological frequency data. We will use the name *Syncretism Distribution Modeling* (SDM) for this technique. We will use SDM to compare different theories of d-partition distributions. We conceive of an analysis of the paradigm patterns as a condition specifying which m-partitions are available as possible parameter settings. For example, one theory of grammar may say that only d-partitions 1, 3, 15 are available as m-partitions. The set of m-partitions available constitutes a *Morphological Parameter Space*. If d-partitions other than those occurring in the paramater space occur, they must arise via accidental homophony. Note that the d-partition with as many morphemes as cells cannot arise via accidental homophony, so if it occurs, the morphological parameter space must contain this paradigm.

Generally when we have 4 cells, any subset of the set of 15 paradigms would constitute a parameter space. But since four different exponents for first person actually occur, only parameter spaces that contain paradigm 1 are possible. Then, the total number of possible parameter spaces is then $2^{14} = 16\,192$.

The two parameter spaces that are most important for us in this paper, are the parameter spaces that make the weakest assumptions about universal grammar: Analysis *All-Parameters*, the parameter space that contains all 15 possible m-partitions, and Analysis *All-Accidental*, the parameter space that contains only m-partition 1. Analysis All-Parameters says that all 15 patterns are available as parametric choices. So this analysis assumes that languages can freely vary as to which set of concepts they verbalize in the person paradigm. As far as a universal theory of language, it assumes that part of the description of any language is a set of concepts it verbalizes. Analysis All-Parameters can therefore account for any syncretism as systematic homophony, but it still allows ac-

cidental homophony to occur as well. At the opposite extreme, Analysis All-Accidental assumes that only the paradigm with four distinct cells is available as a parametric choice – as explained above, at least this parametric choice must be available because otherwise Ilocano would not be a possible language. The All-Accidental analysis assume that underlying all languages are like Ilocano and that all syncretism is accidental homophony. As far as universal grammar is concerned, the assumptions of the All-Accidental analysis are even weaker than the All-Parameter analysis. Namely, the All-Accidental analysis assumes that the semantic categories are given independent of language and learning a language only requires acquiring the sounds for each of the semantic categories. In a sense, the All-Parameters and All-Accidental analyses are opposites, but both constitute reasonable null hypotheses. Most active morphologists reject both analyses. But, as we discussed in Section 1.1 the argument against either of the two have rested on problematic assumptions about accidental homophony. The main goal of the present paper is to present an new statistical argument against both null hypotheses. Namely, we argue that both null hypotheses fall short of providing an explanation of the cross-linguistic frequencies of paradigm patterns because they predict the actual pattern frequencies to have an extremely low probability. This reasoning is based on precise probabilistic assumptions about accidental homophony that make it possible to compute the likelihoods of the actual pattern frequencies for different analyses.

3.2 The Probability of Patterns

In this section, we introduce how to compute, for an m-partition, the likelihood of d-partitions that can arise from it given a rate of accidental homophony h. The rate h as we introduced above refers to the likelihood of homophony of two distinct morphemes of an m-partition. The rate h therefore determines for any underlying m-partition m (with #m morphemes) how likely any d-partition e that is at least as coarse as m is to arise. The simplest case is a two-morpheme paradigm. In this case, the application of h is straightforward: the likelihood of the d-partition with just one exponent is the likelihood of the two morphemes being homophonous, i.e. h. By complementarity, the d-partition with two non-homophonous morphemes has likelihood $1-h$.

Now consider paradigms with more than two morphemes. The likelihood of a particular exponent structure arises from a repeated random event, deciding for each morpheme which other morpheme it is homophonous with if it is accidentally homophonous with another morpheme at all. In probability theory such repeated random events are often conceptualized as drawing balls of different colors from an urn. An important factor is whether the ball is returned to the urn or not after any drawing of a ball. If the ball is not returned, the likelihoods in the second drawing depend on the color of the initial ball, and at some point we would run out of balls of one color altogether. For these reasons, the model without return doesn't fit our assumptions: e.g. if cells 1 and 2 are already homophonous, the likelihood of cell 3 to be homophonous with 1 and 2 is still h by our assumption that h is a constant for any two cells. Also a model without return of the balls predicts that at some point a particular color is used up. But that would predict that there must be a

specific number n such any exponent can only be ambiguous in at most n many ways. Such a numerical limit on homophony is implausible for morphology. In sum we conclude that the random event we are looking at is appropriately viewed as an urn model with return of the balls. In such a model, the likelihood of drawing a ball of the same color stays h even after multiple drawings. There is one further caveat: The urn model initially seems to only allow fractions likelihoods: if the urn contains n balls in total, and m of them are of a specific color, then the likelihood of drawing a ball of that color is m/n. We can, however, assume h to vary without a restriction to fractions since any value of h can be approximated to any degree of precision by a fraction.

With these assumptions in place, it is now clear how we can calculate the likelihood of a particular paradigm pattern for a paradigm with m-many morphemes. The general formula for a paradigm with m morphemes and e distinct exponents is given in (2) (assuming $0 \leq h \leq 1/(m-1)$):

$$\Lambda_h^m(e) = h^{m-e} \prod_{i=1}^{e-1} (1 - ih) \tag{2}$$

In the case of $m = 2$, the formula in (2) says that the d-partition with $e = 2$ distinct exponents has likelihood $1 - h$ and the one with only one exponent ($e = 1$) has likelihood s. If $m = 3$, we can build upon the case of two cells. If cell 1 and cell 2 are homophonous, the chance of cell 3 also being homophonous is also h. So overall the chance of accidental homophony of all three cells is h^2, while the chance of homophony of 1 and 2, but excluding 3 is $h(1-h)$. Now consider the case that cell 1 and cell 2 aren't homophonous. Cell 3 then has chance h of being homophonous to cell 1, chance h of being homophonous to cell 2 and $1 - 2h$ chance of not being homophonous to cell 1 or 2.

Formula (2) entails that the likelihood of two paradigm patterns is the same if they involve the same number of exponents, regardless of the number of times each exponent occurs. We want to show by means of an example that this is correct: With four cells, there are two different possibilities to fill the paradigm with two exponents, E_1 and E_2: either E_1 occurs three times and E_2 only once, or E_1 and E_2 both occur twice. Concretely consider the likelihood of the two patterns $\langle \mathrm{E}_1, \mathrm{E}_1, \mathrm{E}_1, \mathrm{E}_2 \rangle$ and $\langle \mathrm{E}_1, \mathrm{E}_1, \mathrm{E}_2, \mathrm{E}_2 \rangle$. In the first case, morpheme 2 and morpheme 3 both have likelihood h to be identical to morpheme 1, and morpheme 4 has likelihood $1 - h$. In total then the first pattern has likelihood $h^2(1-h)$. In the second pattern, the likelihood of morpheme 2 being identical to morpheme 1 is h, and the likelihood of morpheme 4 and morpheme 3 being identical is also h. Furthermore the likelihood of morpheme 1 and morpheme 3 being distinct is $1 - h$, so we also arrive at $h^2(1-h)$ as the likelihood for the second pattern.

Finally, consider our assumption that h be no greater than $1/(m-1)$ that we already mentioned above. The reason for this assumption is that for most h greater than $1/(m-1)$, our basic assumption that h is constant for any selection of two distinct morphemes from an m-partition leads to an inconsistency. Consider the case of 4 morphemes and $h = 40\%$. Now assume that, in one specific language, morphemes 1, 2, and 3 are all different. The likelihood of morpheme 4 to then be accidentally homophonous to any of 1, 2, and 3 must

be equal to h. But, these three likelihoods of mutually exclusive events add up to 120%. This contradiction shows that h cannot be as big as 40%.[5] Note in addition that the formula as stated above would yield negative values in some of the cases ruled out.

3.3 Calculating the Best Fit Syncretism Rate for Model 1

In this section, we show how to find h so that it best fits a set of data in the case of a model where just one parameter choice exists, i.e. what we termed above *All-Accidental* models. For the case of the four first person cells, we call this model *Model 1*. The approach developed with just one parametric choice can then be extended for models with more parametric choices by adding up the prediction for each parameter choice. The kind of predictions our approach makes are statistical unless $h = 0$. This means that generally any frequency distribution for the d-partitions is possible, but crucially the likelihoods of different distributions will differ. The general technique we will then apply to compute a value for h is maximum likelihood modeling (see also section 3.6).

So our immediate goal is to find the value of h such that it makes the observed distribution maximally likely for the given parameter space. Concretely, we consider Model 1 containing a single m-partition. In other words, we assume that the morpheme structure of all 83 paradigms is the one of Ilocano: four different morphemes. For any value of h, the expected frequency of d-partition e is given by $83\Lambda_h^4(\#e)$ where $\#e$ is the number of exponents of d-partition e. Consider for example $h = 0.1$: The relevant coefficients are $\Lambda_h^4(1) = h^3 = 0.1\%$, $\Lambda_h^4(2) = h^2(1-h) = 0.9\%$, $\Lambda_h^4(3) = h(1-h)(1-2h) = 7.2\%$, and $\Lambda_h^4(3) = (1-h)(1-2h)(1-3h) = 50.4\%$. From these we can compute the expected distribution when there are 83 total occurrences: 0.083 occurrences are expected for d-partition 15 with one exponent, where actually 15 are observed. We also expect 0.75 occurrences each for d-partitions 8 through 14 with two exponents each and 6.0 occurrences for each of the three-exponent d-partitions 2 through 7. Finally, 41.8 occurrences of the single four-exponent d-partition 1 are expected. The expected frequency distribution is, of course, not predicted as the actual frequency distribution, but rather is the mean point of a probability distribution. I.e. if we repeatedly selected 83 languages at random and took the mean number of occurrences across these trials, the means ought to converge against this number.

The h with the best fit between the expected distribution and the observed distribution is determined by the distance between the two. Both the expected distribution and the observed distribution are vectors of 15 values that define points in a 15 dimensional space. One way to do this would be use the Euclidean distance of the two points, i.e. the square root of $\sum_e (O_e - E_e^h)^2$, where E_e^h are the expected (depending on h) and O_e the observed frequency of d-partition e. But using Euclidean distances would be sensitive only to absolute differences, whereas intuitively relative differences are also important. Consider, for example, widely different relative deviations between E and O: assume in one component $E_e = 1$ and $O_e = 50$ while in the other $E_{e'} = 50$ and $O_{e'} = 100$. With Euclidean distances the latter difference would still count as worse than the former, even though in the former

[5] For $h = 1/n$ for any integer n, no problem arises even if $h > 1/(m-1)$.

O_e is off by factor 50. Therefore the use of Euclidean distances would lead us to have to accept much larger relative deviations between E and O when O and E are both small than when they are large. This outcome is generally regarded as undesirable. Instead we will use the following measure of distance: $\text{fit}(h) = \sum_e (O_e - E_e^h)^2 / E_e^h$. While we can not fully justify this decision at this point of the paper, note that this is the kind of measure that the well-known chi-square statistic makes use of.[6] In section 3.6, it will become clear that this distance measure ensures the h is chosen so that it maximizes the probability of an event of the type of O. The solid line in figure 5 shows fit between expected and the observed distribution for h within the interval $[0.15,\ 0.33]$. We used the function *optimize* of the statistics software R to find the minimum of fit automatically: The accidental homophony rate that gives the closest fit is 28.55%.

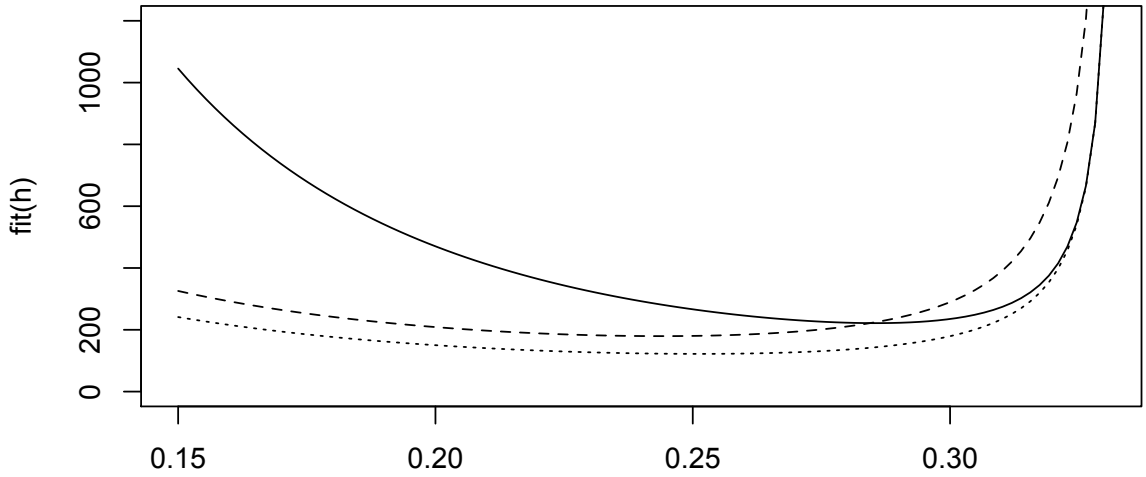

Figure 5: Dependence of fit on the rate of accidental homophony for Models 1 (solid line), 2 (dashed), and 2′ (dotted).

In sum, we have shown how different rates of accidental homophony lead us to expect a different frequencies of the d-partitions even when all languages share the same m-partition. We can then compute the fit between the expected distribution and an observed distribution and use this to choose the accidental homophony rate h so that the fit is best. However, in this section we only considered a model where all languages have the same m-partition. The following section shows how we can compute the expected distribution and again find the h with the best fit.

[6]Other statistical procedures rely on other computations of distance. For example, the G-test relies on log-likelihood ratios. We opt for the chi-square based computation at this point because the statistics software R we use has a fuller implementation of the chi-square test, while we will need to implement a Monte Carlo versions of other tests ourselves for the computations in Section 3.6.

3.4 Parameter Choice and Model 2

Linguists have used the term parameter to describe linguistic variation (Chomsky and Lasnik 1993). A parameter space captures the space of possible languages, and each language is associated with a parameter choice.[7] In Model 1 there was no meaningful parameter choice – all languages shared the same m-partition. We now introduce Model 2 that allows a parameter choice: Namely it allows a language to be specified for either m-partition 1 (Ilocano) or m-partition 13 (English). Then, the expected distribution E^h depends not only on the likelihood of accidental homophony, but also on the likelihoods of the two parameter choices. The likelihoods of parameter choices are difficult to determine *a priori* at this point,[8] so we limit our attention to two approaches: On the one hand, an equal distribution (*Equi-Parametrization*) where both m-partitions have likelihood 50%. And on the other hand, a best fit distribution (*Best Fit Parametrization*) where the likelihood of the m-partitions is optimized to result in the best to an observed distribution. Specifically, we introduce *Model 2* as an equi-parametrization model in the first person case and *Model* $2'$ as a best-fit parametrization version of model 2. In this subsection, we focus on the equi-parametrization and model 2, and return to the best-fit parameterization and model $2'$ in the next section.

The equal distribution generally assumes that each parameter choice has the same likelihood. For example, if 5 m-partitions are available as parameter choices, each is assumed to have likelihood 20%. For Model 2, however, each of the two m-partitions has likelihood 50%. In the absence of considerations arguing against equal likelihoods, we think the equal distribution is the best starting point. Furthermore in at least one domain of parametrization, an equal distribution is attested: Gilligan (1987) provides data on agreement from a biased sample of 100 languages. He considers which arguments a verb agrees with. There are four parametric choices in this case: no agreement at all, agreement with the subject, agreement with subject and direct object, and agreement with subject, direct object and indirect object. Gilligan's findings are reproduced in table below[9] The G-test (Sokal and Rohlf 1995) shows that Gilligan's distribution is not significantly different from an equal distribution ($G(3) = 2.1785$, p-value$= 0.5362$).

(3)

parameterization	no Agreement	S only	S and DO	S, IO and DO
# of languages	23	21	31	25

Once the likelihoods of the parameterizations are determined, we can compute the overall expected distribution by computing the expected distribution for each possible parameter choice given the percentage of language with that parameter choice and then adding up the

[7] Or possible a likelihood profile for the possible parameter choices. (Yang 2002)

[8] Future work may be able to determine a priori rates, for example by the application of game-theoretic methods.

[9] Gilligan actually reports the language Waskia as having agreement with the subject and the indirect object. However, Bobaljik (2008b) argues that Waskia actually has a form of suppletion with *give*, not indirect object agreement. Hence we count Waskia as a language with only subject agreement.

results from the individual parameter choices. In the case of model 2, we show below that this is straightforward. In the general case, assume that the parameter space contains the m-partitions $m_2, \ldots, m_n$ in addition to m_1, the complete m-partition space which is needed for the analysis of languages without syncretism. We again use $\#m$ for the number of morphemes in m-partition m. Note that $\#m_1 = n$, but $\#m_i < n$ for $i > 1$. Assume that the vector $\vec{a}$ contains the likelihood for each parametrization, i.e. a_i be the likelihood that a language will have parametrization m_i for each i. Since m_1 through m_n are by assumption all the available parameterizations, $\sum \vec{a} = 1$. In the case of the equi-parametrization, $a_i = 1/n$ for any $i \in \{1, \ldots, n\}$.

If $h = 0$, the expected distribution E^h equals $\ell\vec{a}$, where ℓ is the number of paradigms sampled. But if $h > 0$, the distribution of d-partitions of the languages with different m-partitions can be affected differently depending depending on which d-partitions are courser than the m-partition and on the number of morphemes in an m-partition. The former effect is due to the fact that from each m-partition only coarser d-partitions can arise via accidental homophony. The later is the observation that, at the same general rate h, m-partitions with more morphemes are more likely to exhibit accidental homophony than m-partitions with fewer morphemes. Consider for example Model 2 with $h = 0.1$: Since parameter choice 2 has only two morphemes while choice 1 has four, the expected rate of d-partitions without accidental homophony is 90% of the language with parameter choice 2, but only 50.4% of the languages with choice 1. Let d be a d-partitions. If d is at least as coarse as the m-partition m of a parameter choice, the Λ_h coefficient introduced above gives the likelihood of d-partition d. As argued above, the coefficient actually depends only on the number of morphemes in m, $\#m$, and the number of exponents in d, $\#d$. Furthermore, the likelihood of arising from m is 0 regardless of h for any d-partition other than those at least as coarse as m. Therefore, we can define a matrix $M(h)$ that specifies for each m-partition of the parameters space m and for each d-partition d, the likelihood of d arising for a language with parametrization m:

$$E^h_{i,j} = M^h_{d_i,m_j} = \begin{cases} \Lambda_h^{\#m_j}(\#d_i) & \text{if } d_i \text{ is at least as coarse as } m_j \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

The number of rows of matrix $M(h)$ is n, the number of m-partitions in the parameter space. The number of columns is the number of possible d-partitions, i.e. B_c, the Bell number for the number of cells in the paradigm under consideration.

For any expected distribution $\vec{a}$, the expected distribution of d-partitions can now be computed by multiplication with the matrix M and the number of languages/paradigms in the sample ℓ:

$$\ell M \times \vec{a} = \langle \sum_{i=1}^{n} M_{e_1,m_i} a_m, \sum_{i=1}^{n} M_{e_1,m} a_m, \ldots, \sum_{i=1}^{n} M_{e_{B_c},m} a_m \rangle \tag{5}$$

The equal distribution is a specific application of result (5), namely the case of $a_i = 1/n$ for all $i \in \{1, \ldots, n\}$.

For illustration, consider again Model 2: Then the matrix M^h has two rows and the following values where the second row contains only 0 except for columns 13 and 15:

$$\begin{pmatrix} (1-h)(1-2h)(1-3h) & h(1-h)(1-2h) & \cdots & h^2(1-h) & \cdots & h^2(1-h) & h^2(1-h) & h^3 \\ 0 & 0 & \cdots & 0 & \cdots & 1-s & 0 & s \end{pmatrix}$$

Finding the best fitting h for the equi-parametrization now requires minimizing the following term for M^h as just given above (recall that ℓ is the number of paradigms sampled):

$$\text{fit}(h) = \left| \ell M^h \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix} - O \right|^2 \tag{6}$$

One main effect of adding m_2 to the parameter space is that because of the equi-parameterization only 50% of the samples are assumed to have m-partition 1, rather than 100%. For low values of h, we expect model 2 therefore to be better than model 1. For high values h, however, the expected frequency of paradigm 1 is going to be less than the observed, and here model 1 will have an advantage. These intuitions are confirmed by our results: The dashed line in figure 5 shows how h affects the fit of model 2. Using again the optimize-function of R, we determine that the best fitting accidental homophony rate h for model 2 is 24.39%. The best fit made possible by model 2 is better than that of model 1. However, we will see below that model 2 still needs to be rejected.

3.5 Best Fit Parametrization and Model 2′

The second possibility we consider is the best fit distribution where we determine the assumed distribution from the actual data so as to minimize the difference between the expected and observed distribution. Though the equal distribution is preferred and our investigation in the following is going to mostly use it, the best fit distribution is a useful comparison. Consider first the problems of the best fit parametrization, and then the ways it can be useful. The best fit distribution raises the problem of overfitting the data: the resulting morphological models have two parameters: h and the distribution of parameterizations, which is itself a vector whose length k is equal to the number of possible m-partitions and has $k-1$ degrees of freedom. Since our investigation rests on only one data point – the actual distribution of languages – and there cannot be more than this one data point, the number of variables of the best fit model is likely to make the models too permissive. However, the best fit model is useful when it rejects a morphological system: Clearly if even the overly permissive best fit parametrization doesn't allow us to fit the actual e-distribution well, a morphological parameter space can safely be rejected.

The mathematical computation involved in the best fit parametrization is more complex than the equal distribution models. Furthermore the result of the optimization depends on the accidental homophony rate h. We consider h fixed in the following, though, because we can actually find the best fit parametrization for each h within the minimization of h.

The optimization problem is then specified by the term in (7). Because the likelihoods of all paradigms possible must add up to 1, one component of the distribution vector $\vec{a}$ is determined by the other $n-1$ components. Here we assume that the last component is the dependent one.

$$\text{fit}(p_1,\ldots,p_{n-1}) = \left| \ell M(h) \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_{n-1} \\ \ell - \sum_{i=1}^{n-1} p_i \end{pmatrix} - O \right|^2 \tag{7}$$

There is an additional constraint on $\vec{p}$ that needs to be incorporated. Namely, none of its components may be negative since they represent probabilities. But it is frequently the case that the minimum of (7) has negative components: If the d-partitions in position i occurs not at all or very rarely but is available as m-partitions, a negative value of p_i in a sense would counterbalance the positive contribution to the i-th d-partition made by a frequent m-partition may via accidental homophony. To rule out negative solutions, the values $p_1,\ldots,p_{n-1}$ must each be greater or equal to 0, and furthermore $\ell \geq \sum_{i=1}^{n-1} p_i \geq 0$ must be greater or equal to 0. The following condition restates these requirements as an equation in the form $Ap+b \geq 0$ as required by the function *constrOptim* of the software R.

$$\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -1 & -1 & -1 & \cdots & -1 \end{pmatrix} * \vec{p} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ -\ell \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{8}$$

Finally, we need to determine the gradient of the function *fit*. The gradient is useful to find the minimum of a function quickly with numerical algorithms. In general, the gradient ∇f of a function f taking k-long vectors as arguments is a function returning also k-long vectors that describe the slope of the surface f defined by f in a $k+1$ dimensional space. Intuitively speaking, the gradient is useful to find the minimum because local minima have slope 0 and in other points the slope gives us information in which direction a local minimum can be found. In the present case, the function *fit* may have many components and therefore it would be computationally very inefficient to search minima without knowing the gradient. To compute the gradient, note that *fit* is the concatenation of the quadratic function $f(x) = x^2$, the linear function $g(x) = Mx - O$ and the function $h(p_1,\ldots,p_{n-1}) = (p_1,\ldots,p_{n-1},1-\sum_{i=1}^{n-1} p_i$. Since the derivative of $(f \circ g)'(x) = f'(g(x))g'(x)$, the gradient can be easily determined as the following, where the $n \times (n-1)$ matrix with all 1 on the diagonal, all -1 on the last row and 0 everywhere else is the derivative of h:

$$(9)\quad \nabla\text{fit}(p_1,\ldots,p_{n-1}) = 2(M\begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_{n-1} \\ 1-\sum_{i=1}^{n-1} p_i \end{pmatrix} - O)M\begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \\ -1 & -1 & \cdots & -1 & -1 \end{pmatrix}$$

As an example of a best fit parametrization. consider model $2'$ introduced above. For model 1 evidently the best-fit parametrization would not differ from the equi-parameterization since there is only one parameter choice. For model $2'$, however, we expect the best fit parametrization to balance the effects of the two possible m-partitions. For the discussion, recall that a_1 be the parametrization for m-partition 1. Looking only at d-partitions 1 and 13, m-partition 1 requires h to be such that $a_1(1-h)(1-2h)(1-3h)$ is close to 18%, the percentage of the observed data with d-partition 1. From m-partition 2 on the other hand, $(1-a_1)(1-h)$ should also be close to 18%. Intuitively model $2'$ should improve on model 2 if h is reduced from 27%, and a_1 is increased above 50% because the the contribution of m-partition 1 to d-partition 1 depends negatively on p^3. For example, with $h = 20\%$, the two expressions $a_1(1-h)(1-2h)(1-3h) = 0.192a_1$ and $(1-a_1)(1-h) = 0.8-0.8a_1$ are equal for $p = 0.806$. So if $h = 20\%$, the two approximation requirements mentioned above would be best when about 80% of the paradigms are assigned to m-partition 1. When we take all 15 d-partitions into account in a numerical calculation, the best-fit values turn out to be $h = 19.5\%$ and $p = 76.0\%$ for model $2'$. The dot line in figure 5 shows how for different h values, the fit of the best-fit parametrization of model $2'$ compares to the fit of model 1 and model 2 with the equi-distributed parametrization. For logical reasons, the fit of Model $2'$ cannot be worse than Model 1 or Model 2 since both of them are special parametrizations of $2'$. Actually, the best-fit parametrization generates a better fit than either Model 1 or 2. However, at this point we can't yet say what would constitute a satisfactory fit. In the next section, we address this issue.

3.6 The Likelihood of the Observed

In this section, we develop the method to evaluate the approximizations introduced in the previous sections in theory independent terms. The starting point is to recall that the result of the computations introduced above is in each case a probability distribution. We have presented methods to adjust the syncretism rate h so that the expected distribution of languages is as close as possible to the observed distribution. In this section, we discuss how to determine what likelihood a model assigns to an event like the observed distribution. Before we introduce that in detail, we discuss how to compare models based on such a likelihood assigned to the observation.

The line of statistical reasoning we employ is called *Maximum Likelihood Modeling*. The perspective is substantially different from the statistics of hypothesis testing that

is used in psycholinguistic studies. In psycholinguistics, the focus is usually on showing that a null hypothesis can be ruled out. Typically, a psycholinguist aims at a demonstration that one variable that varies across the set of observations considered predicts the experimental data better than is predicted by the null hypothesis—most often the null hypothesis is simply that the variable shouldn't affect the observations at all. To this end, a psycholinguist computes, for the model with the null hypothesis, how likely an outcome such as the observed outcome was, and rules against the null hypothesis if the computed p-value is low. If the null hypothesis is rejected, this then usually argues for an alternative hypothesis where the variable should affect the observation. Since the ideal experimental design used is one that is precisely tailored to only distinguish the null hypothesis from the alternative hypothesis, usually the alternative hypothesis is corroborated whenever the null hypothesis can be ruled out.

This type of reasoning also plays a role in our investigation, but a more limited role than is typically the case. Specifically, we want to rule out the two quasi null hypotheses we introduced above: the All-Accidental and the All-Parameters model. However, the data from syncretism pattern frequencies are much richer than the data gathered in a typical psycholinguistic experiment. The real challenge turns out to be to find a principle based hypothesis that is not ruled out by the observed data. But to accept an alternative hypothesis, the p-value for it should be high. So, while low p-values for the null hypotheses support a model that diverges from it, that model itself then also needs to have a high p-value since otherwise we would also feel compelled to rule out the alternative hypothesis. A situation, that the alternative hypothesis as well as the null hypothesis is ruled out by the observations can also occur in psycholinguistic experiment. For example, a predicted difference may go 'in the wrong direction' as follows: assume the alternative hypothesis predict that one set of observations should be greater than another set and the null hypothesis predict them to be not different. Then if we actually observe that the first set is on average smaller than the second set of observations, this is not consistent with either hypothesis. In actual psycholinguistic work, however, such an outcome seems to occur rarely as far as we know.

Both types of reasoning require the computation of the probability of a set of observation given a model. In our case, the observation is the observed distribution of patterns in Cysouw's sample. Of course, though the probability of any single datapoint is always vanishingly small. The interesting probability is not only the one of this single point, but the p-value: that of all points that are at least as far away from the expected distribution as the datapoint under consideration – in our case, the datapoint from Cysouw's sample. The reasoning is transparent in the example of the dice count we mentioned already in the introduction: imagine you don't know how many dice I rolled, but you know that I rolled a 139. Does this allow you to reject the assumption that I rolled 40 dice (i.e. the 40-dice model)? The likelihood of rolling a 139 with 40 dice is about .036, i.e. below the .05 level. But clearly that should not lead to rejecting the 40-dice model: actually 140 is the most likely expected result when rolling 40 dice, and 139 is very close to that. The likelihood of being 1 or more away from 140 is greater than .96, and for this reason we certainly cannot reject the 40-dice model. On the other hand, the assumption that I rolled 139 dice all showing a one, the 139-dice model, can be rejected, because only the event of rolling

a $6 \times 139 = 834$ is at least as extreme as 139 on the 139-dice model, so the p-value of 139, i.e. the likelihood of an outcome as extreme as 139, is smaller than 10^{-100} (namely, $2 \times 1/6^{139}$) on the 139-dice model.

What is an acceptable p-value for a model in maximum likelihood modeling? For hypothesis testing, the .05 level is established in the psycholinguistic community as the threshold that a p-value should be below in a published paper, though in some areas of neuroscience much lower p-values like $p < .001$ are standard. As far as we know, no corresponding threshold is established for maximum likelihood modeling. The .05 level is relevant: If a model led to a p-value greater than .05, we would not be licensed to reject that model as inconsistent with our data. So this would be one outcome worth reporting. But there is another outcome worth reporting, which involves a relative improvement in p-value from a null hypothesis. For example, while a null hypothesis may predict the observed data to be close to impossible (e.g. a p-value of 10^{-20}), an alternative hypothesis may predict a much higher likelihood for the observed data (e.g. 10^{-3}). In this case, the enormous relative improvement in accounting for the observed data should be recognized as an important step towards an explanation of the observed data, though the .05-level is not yet reached. Therefore relative improvement of the p-values is as important as any absolute level. In the case of syncretism patterns specifically, it would be surprising if one alternative hypothesis about parametrization alone would completely explain the data since as we mentioned above the independence assumptions may be to some extent idealizations and the data may be difficult to model because of that.

In the remainder of the section, we now address the appropriate way to compute a p-value in our scenario. Unless h is 0, the result of each model is a probability distribution where each d-partition is predicted to occur with some likelihood greater than 0. The observed outcome for the four cells of first person was shown above, and in other cases it will be similar: a distribution where some d-partitions occur a sizable number of times, possibly even greater the cap of 15 we applied above, while others occur very rarely or not at all. Two widely used tests for computing how likely an observed distribution across a number of cells is given a probability distribution: the Chi-Square test and the G-test.[10] These tests provide a p-value for an observed distribution assuming a random event that can have finitely many discrete outcomes (or *bins*). In our case, the bins are the d-partitions. However, both cannot be used when the observed distribution contains 0s or even just counts below 5 for some bins.[11] As we saw above that even with the full data available to us from 265 paradigms, only 5 of the 15 bins of in the first person case have count of 5 or greater. One might want to address this by gathering more data, however, this would require a massive effort in the case of eight cells, which holds the greatest interest: With eight cells, the number of bins is $B_8 = 4140$. In principle this might be possible, there probably are many more than the roughly 6000 languages recognized by Grimes (2000) once the independent status of so-called dialects is recognized, and in addition many languages

[10] In the future, we plan to also apply the root-mean-square test of Perkins et al. (2011).

[11] One way to address this problem would be to merge some of the bins so that all observed frequencies are at least 5. However, we have not explored this yet. For one, the results are strongly affected by the rebinning chosen. Furthermore, in a case with many 0s much of the power of the test is lost.

have more than one relevant paradigm. Still, to gather a sample where all of over 4000 cells occurred at least five times is far beyond our current capabilities.

Since there is, as far as we presently know, no other statistical test applicable in our scenario, we are left with what is called in statistics the Monte Carlo method. This means we generate a random event that is distributed as our model predicts and count in a sample output how often an event at least as extreme as the real observation occurs. In the statistics software package R, the *chisq.test* defaults to such a simulation when given an input that doesn't satisfy the requirement that all cells contain a count of at least 5 and we use this function in the end to compute the p-values for our data. In fact this choice already affected our choice of optimization procedure above: we minimized the term $\sum(E-O)^2/E$ because this is the term the chi-square test depends on. For future research it is important to keep in mind that the distance metric used above needs to be chosen in accordance to the statistic test used at this point.

The Monte-Carlo method is slow method to estimate a likelihood, and this in particular affects very low likelihoods. The Monte-Carlo method computes a large number of random events and counts how many of these are as far away from the expected distribution as the observed one. For a high probability this provides a useful estimate: for example, if 3500 of one million trials turn out to count, then 0.0035 is highly likely to be good estimate of the p-value. For very low likelihoods, it is interesting to know whether they are 10^{-10} or 10^{-20}. But with only one million trials, we are not able to distinguish such low likelihoods because for neither one are likely to encounter even one event that counts. All we could conclude from one million trials in this case, would be that the likelihood is below 10^{-6}, i.e. the likelihood of one in a million. At this point, our computational resources don't allow us to be more accurate than this. All we can say is that all the three models considered so far assign a very low likelihood to the observed distribution of languages O. The computation of two million trials underlying each line of the following table requires about 10 minutes per model on an Apple iMac personal computer with a 3.06 GHz Intel Core 2 Duo microprocessor.

model	h	fit(h)	p-value of O
Model 1	0.2855	221.5	$< 10^{-6}$
Model 2	0.2439	179.5	$< 10^{-6}$
Model 2′	.2512	122.0	$< 10^{-6}$

The low p-values that all three models assign the observed distribution entail that all three models can be rejected. Looking at the fit(h) in the table and also in figure 5, we see that model 2 is better than model 1. But we can't compute the improvement of model 2 relative to model 1 in terms of p-value: even model 2′ where the parametrization is made to best-fit the observed data still assigns to the observed distribution a likelihood below one in a million. The rejection of model 1 is particularly interesting since it is the instantiation of the all-accidental model in this case, i.e. the general model Cysouw (2003) advocates.

Several further models are interesting and at this point everything is in place to just test them. We list the relevant values for four more models in the following table,

where the Monte Carlo testing was only done with 20 000 repetitions. Model $\forall$ is the all-parameters models. This model, as we mentioned, can also be considered a null hypothesis that makes only very minimal assumptions about grammar. The fact that it just like Model 1 can be safely rejected, shows that universal grammar constrains the parameter space of m-partitions in a non-trivial way.

model	h	fit(h)	p-value of O
Model $\forall$	.012	122.7	$< 10^{-4}$
Model > 14	.096	18.3	=.190
Model > 2	.055	15.3	=.106
Model > 1	.041	24.0	=.009

Models > 14, > 2, and > 1 all compute the parameter space directly from the observed distribution in figure 4. Namely, the parameter space consists of the m-partitions that as d-partitions have an observed frequency above the threshold of 14, 2, and 1 respectively. We expect these models to assign a high likelihood to the actual distribution, and this is in fact the case. This shows that some analyses that combine a parameter space restricted by universal grammar and some accidental homophony are consistent with Cysouw's dataset.

4 Conclusions and Outlook

The distribution of patterns of identity and non-identity among paradigm cells (our d-partitions) across the world's languages is unquestionably not random. Yet accidental homophony introduces an element of noise into these distributions, making it hard to distinguish the accidental from the systematic. In this paper, we have proposed a new approach to this problem, using statistical techniques: Syncretism Distribution Modeling.

Our main thesis extends a familiar premise using more precise mathematics: We treat accidental homophony as a random event in the statistical sense. Thus the difference between the systematic and the random should emerges in the formal analysis of large datasets. In section 3, we showed how to compute the p-value of an observed distribution of (d-)partitions of the paradigm space, assuming a given morphological model M. The p-value as computed in the 3.6 provides us with a direct way to interpret the fit between model M and observation O: Assuming model M, an outcome at least as extreme as observation O has likelihood p of occurring. By iterated application of this procedure, we can compare different morphological models by their fit with an observed cross-linguistic distribution of d-partitions. In this situation, we can then apply maximum likelihood modeling to determine the model or models that assign the highest likelihood to the actual observation O.

For reasons of space and time, in this paper, we described only the initial steps in this project, offering in essence a proof of principle. We focussed on the case of person paradigms, and at that mostly on the sub-paradigm created by the four cells of the first person, using an existing dataset (Cysouw 2001). The main linguistic result we report here is that analyses that assume a parameter space unconstrained by linguistic considerations (the

models 1 and ∀ above) cannot explain the observed frequencies of d-partitions in this case. In ongoing work, we supplement these methods with an automatic procedure to compute a parameter space from a set of features. This gives us a way to directly assess the goodness of fit of different feature sets. For example, we may successfully infer a particular model (i.e. universal feature inventory) provides a good fit for the observed data – in fact one that converges with semantically plausible models assumed in the theoretical morphological literature – using this methodology. In addition we can compare the effect of allowing explicit ordering of the features or only Paninian implicit order. Our ongoing work addresses a question that we can only hint at here: in what way the model should be constrained? Here we pointed out that some parameter spaces (Model > 14 and Model > 2 above) fit the observed distribution well, but this is clearly a post-hoc result. The level of parameter spaces is for two reasons not the right level to look for satisfactory morphological models. For one, there are over 16 000 potential parameter spaces for the case of 4 cell paradigms, and given this large number, it is less surprising that some of them fit the observed distribution. Secondly no theory of grammar that we know of assumes that the parameter space is specified as a set of m-partitions. Rather the set of m-partitions that characterize a parameter space are only a derivative concept and the underlying assumptions are about features and the possible ways features can be related to morphemes.

A further step to be taken is to scale syncretism distribution modeling up to examine the full eight cell paradigm. While the principles are identical, scaling up turns out to be computationally challenging and might require more powerful computers than we can presently access. We conclude by noting that other four cell selections from the full eight cell paradigm exhibit similarly marked distributions as the first person selection that we focussed on. The d-partition counts of three other selections are shown in figure 6. For each, the pattern in the top left corner of each diagram shows by shading which cells were selected. The leftmost diagram shows the counts for the four non-first person cells. The center concerns the minimal cells, and the right shows a selection of cells that don't have any natural semantic category corresponding to them. The diagrams show that the person distribution is unusual also outside the first person domain.

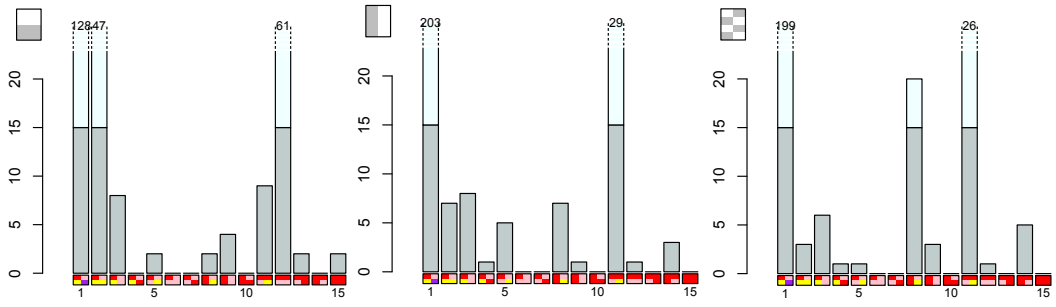

Figure 6: frequencies of d-partitions for three four cell selections other than first person

References

Anderson, Stephen R. 1992. *A-morphous morphology*. Cambridge University Press.

Berko, Jean. 1958. The child's learning of English morphology. *Word* 14:150–177.

Bobaljik, Jonathan D. 2008a. Missing persons: A case study in morphological universals. *The Linguistic Review* 25:203–230.

Bobaljik, Jonathan D. 2008b. Where's ϕ? agreement as a post-syntactic operation. In *Phi theory: Phi-features across modules and interfaces*, ed. David Adger, Daniel Harbour, and Susana Béjar, 295–328. Oxford, UK: Oxford University Press.

Chomsky, Noam, and Howard Lasnik. 1993. Principles and parameters theory. In *Syntax: Ein internationales Handbuch der zeitgenssischen Forschung*, ed. Joachim Jacobs, Arnim von Stechow, Wolfgang Sternefeld, and Theo Vennemann, 506–569. Berlin, Germany: de Gruyter.

Cysouw, Michael. 2003. *The paradigmatic structure of person marking*. Oxford, UK: Oxford University Press.

Cysouw, Michael. 2011. The expression of person and number: a typologists perspective. *Morphology* 21:419–443.

Cysouw, Michael Alexander. 2001. The paradigmatic structure of person marking. Doctoral Dissertation, Radboud University Nijmegen, Nijmegen, Netherlands.

Forchheimer, Paul. 1953. *The category of person in language*. Berlin, Germany: de Gruyter.

Gilligan, Gary M. 1987. A cross-linguistic approach to the pro-drop parameter. Doctoral Dissertation, University of Southern California.

Grimes, Barbara, ed. 2000. *Ethnologue*. Dallas, Texas: SIL International.

Halle, Morris, and Alec Marantz. 2008. Clarifying blur: Paradigms, defaults and inflectional classes. In *Inflectional identity*, ed. Assif Bachrach and Andrew Nevins. Oxford, UK: Oxford University Press.

Harbour, Daniel. 2008. On homophony and methodology in morphology. *Morphology* 18:75–92.

Harley, Heidi, and Elizabeth Ritter. 2002. A feature-geometric analysis of person and number. *Language* 78:482–526.

Perkins, William, Mark Tygert, and Rachel Ward. 2011. Computing the confidence levels for a root-mean-square test of goodness-of-fit. *Applied Mathematics and Computation* 217:9072–9084.

Pertsova, Katja. 2011. Grounding systematic syncretism in learning. *Linguistic Inquiry* 42:225–266.

Simon, Horst. 2005. Only *you*? philological investigations into the alleged inclusive-exclusive distinction in the secon person plural. In *Clusivity*, ed. Elena Filimonova, 113–150. Amsterdam: John Benjamins.

Sokal, Robert R., and F. James Rohlf. 1995. *Biometry: The principles and practice of statistics in biological research*. New York: Freeman, 3rd edition.

Stump, Gregory T. 2001. *Inflectional morphology: A theory of paradigm structure*. Cambridge University Press.

Thomas, David. 1955. Three analyses of the Ilocano pronoun system. *Word* 11:204–208.
Wechsler, Stephen. 2010. What 'you' and 'I' mean to each other: Person indexicals, self-ascription, and theory of mind. *Language* 86:332–365.
Yang, Charles D. 2002. *Knowledge and learning in natural language*. Oxford, UK: Oxford University Press.