Contradictions and their use in falsification : the case of comparative linguistics and QCA's contribution

Alain Gottcheiner
Laboratoire de Mathématiques et Sciences Sociales
Université Libre de Bruxelles – CP 135 – Av Roosevelt 50 – B 1050 Bruxelles
agot@ulb.ac.be

Abstract

Linguists searching about laws of phonetic changes make use of the entire corpus at their disposal. By so doing, they find laws that correctly describe observed changes, especially « splits », but can’t be checked. Such a law may always be found if using enough parameters, but doesn't guarantee a fair description. In a Popperian perspective, we’d like to suggest working on a partial corpus, trying to establish laws that correctly account for all matching multiplets considered, then applying these assumed laws to the rest of the corpus ; if no counterexample is found, the set of laws gains in credibility.
In this approach, QCA may be very useful, because it allows us to : 1) consider all possible influences (position in the word, preceding and following phoneme, umlaut/ablaut, position relative to stress, …) as conditions ; 2) use contradictions as guides to the detection of influences we forgot to use ; 3) modify the corpus and set of conditions at will ; 4) produce several laws, among which we may choose the most plausible ; 5) find implications that aren't seen at first glance.

1. The study of phonetic changes

1.1. Comparative linguistics

Throughout this text, the word 'comparative' will be used in a sense quite different from the one specialists of case studies are accustomed with, so let's begin by settling that.
Comparative linguistics is the study of similarities and differences between languages (lexicon, syntax, phonetics and phonology) made in the hope of finding :

a) general laws applicable to all, or most, languages (the 'universals of language') ;
b) genetic relationship, both synchronic and diachronic, between specific languages.

The word 'comparative' is also used by linguists to describe a specific method of studying phonetic changes ; see 1.3.

1.2. Phonetic changes

As time goes by, every language undergoes phonetic changes ; that is, some words are pronounced differently, due to a variety of causes : influence from other languages ; spreading of the language, which makes non-native speakers transform sounds according to their articulatory habits ; erosion due to fluent speech. This may make varieties spoken in different areas evolve into separate languages. The usual threshold used for pretending two different languages are born is that spontaneous comprehension between their speakers is no more possible ; this may lead to some paradoxes (the relation 'speaking the same language' is no longer transitive), which may be solved by the use of fuzzy logic. But, as the difference becomes greater, the fact that we're dealing with different languages can't be gainsaid. Phonetic changes aren't the only ones to be involved in the emergence of new languages, but they're the most conspicuous ones, especially when considering unwritten languages.

As an example of what may happen, look at the evolution from archaic Latin to present Spanish.

From the archaic period to the classical times, some phonetic changes happened ; for example, final [am] became [*ã*], as can be inferred from the study of metrics.

When Latin was spread by Roman legions, they encountered local peoples with a wide array of languages, each of which imparted some of its own characteristics to a local version of Latin. The barbaric invasions of 4th to 6th centuries pushed the differentiation further ; for example, what is nowadays Spain fell under Visigothic domination, and this introduced sounds of Germanic origin (like [β]) into local Latin.

At this point, one might consider that Iberic Latin was far enough from Gallic Latin to be considered different languages. In the centuries that followed, they evolved to produce Old Spanish and Old French.

The Arabian invasions and subsequent contact with Arabic language and culture brought many new words (most notably, many words beginning with *al*, a transcription of the Arabic definite article).

At the end of this process, the standard Spanish language (aka *Castillan*) was settled, and differed in many aspects from its neighbours ; while intercomprehension is perhaps possible, through substantial effort, with Portuguese, and with Catalan (which had less contact with Arabic, but more with neighbouring *langue d'Oc*), Castillan is unquestionably distinct from French, for example.

Remarkably, five centuries of separation between European Spanish and Southern American Spanish, brought there in a military way -not dissimilar from how Latin was spread- and contact of the latter with vernacular American languages (Nawa, Mayan languages, Quechua, Guarani …) weren't enough to make it evolve into (a) new language(s) ; its dominating position could be part of the explanation. In the same amount of time, Latin had become a variety of languages, but the Roman Empire had broken up.

The interesting part is that the study of present differences between Spanish and, say, French may give us relevant information about their evolution, to be cross-checked with data from other sciences to make a faithful historic description. While Spanish has acquired Visigothic sounds, French uses some from Western Germanic, like [y], and many words from the same origin.

1.3. The comparative method

The example above is an ideal one ; Latin and Spanish are written languages, and we may study their evolution by reading and interpreting writings from successive epochs. But in the general case, little or no written proof exists, that would enable us to follow the evolution step by step.

Linguists have devised a technique for the study of linguistic changes, based on the study of languages that coexist presently, or at some specific moment in the past ; that is, a synchronic study can give us sound information about past events.

This technique is at its best when applied to phonetic changes, so I will focalise on those from now on. Note that the words "comparative grammar" are often taken to mean above all "comparative phonetics", as is the case in [8b,8d].

Once again I must stress that this "comparative method" had nothing to do with the "comparative method" as described by Ragin [6].

The basic principle of comparative "grammar" is to look for similar words in some languages, and hypothesise that the similarities mean they're linked in some way. Once you've eliminated *motivated* similarities, e.g. onomatopoeia (English *cuckoo* is similar to Japanese *kuku*, with the same meaning, but it's only because they describe the same sound), and borrowings from one language to another, the remaining similarities, if too numerous to be imputed to pure chance, can only be caused by a common origin. And when you've convinced yourself that two words are linked by a common origin, the dissimilarities between them mean that at least one has evolved, and the challenge is to try and reconstruct this evolution.

Take an easy example : English *apple*, Dutch *appel* and German *Apfel* –all with the same meaning- are IOTTMCO linked. There must have been, in some older language, a word that has evolved into those three.

Note that, when the "mother language" is not known, it can't be reconstructed ; there is no way, short of very dangerous speculation, to guess what the original word could have been. Those assumed original words, written only for the sake of easy reference without any certainty about their pronunciation, are usually distinguished by an asterisk : **apal*. In the case of those three languages, their evolution may be followed along the second millennium from Old English (**OE**), Old Dutch (**OD**) and Old High German (**OHG**), but not earlier. The common original language is believed to have been spoken around year 0.

1.4. The fundamental law of comparative phonetics

When studying series of linked words –*cognates*- one may notice *regular correspondences*, that is, the difference between words in a series is the same as in some other series.

For example, along with the Dutch/German pair *appel / Apfel*, one may notice the pairs *paard / Pferd* (horse), *poort* / *Pforte* (gates), …

We may draw a first conclusion : "often, Dutch [p] corresponds to German [pf]". And, according to the principles mentioned in 1.3, "there must be some sound in some old language, which has evolved into those two for some reason". Now we are able to check our hypothesis against other cases. For example, we may look at Dutch *perzik* (peach) and see that it corresponds to German *Pfirsich*, or *kop / Kopf* (head) – everything's fine.

The fundamental principle of comparative phonetics, known under the name *law of regular changes*, may be stated as follows : *any evolution underwent by a sound from one language to another will be the same in every word where this sound appears in the same context*. That is, whenever a Dutch word containing [p] will be a cognate to a German word, the latter will contain [pf] at the same place. And, if one finds a pair where this correspondence doesn't hold, it must be assumed that there is some difference between this word and others that accounts for the differences in their evolution. Linguists have listed such contextual differences : position of the relevant sound in the word (initial vs medial vs terminal), position relative to the main stress (before vs under vs after), identity of the sounds in contact with it (between vowels, before a stop, etc.), identity of other sounds in the word (preceding or following vowel : *ablaut / umlaut*), etc.

1.5. Saussure's hypothesis

The law of regular changes, used unchanged since the dawn of comparative phonetics, in the early 19th century, is a strong statement, and can't be proven. This should make us suspicious about it. Even more suspicious are the hypotheses coming from its unrestricted use.

The most famous of them was made by Ferdinand de Saussure [7] : on seeing that some sounds from reconstructed Proto-Indo-European (**PIE**) seemed to have evolved differently in identical contexts, making the same sound in e.g. Latin correspond to different sounds in Greek without any apparent explanation, he decided to believe, against all expectations, in the law of regular changes and pretend there should have been in PIE some sounds (most probably three), whose presence or absence created differing contexts, which influenced the sounds around them ; then they would have disappeared from all languages coming from PIE, with the consequence that these contexts can't be seen as differing.

Invoking three undetectable phantoms (thereafter called *laryngeals*) for the mere sake of "saving a law" seemed to de Saussure's contemporaries too far-fetched. A more natural attitude would have been to imagine that the fundamental law was too rigid and to look for another one that could explain more phenomena – in the same way as Special Relativity, then General Relativity, generalised the Newtonian laws of movement when contradicting phenomena were observed.

Shortly after de Saussure's death in 1913, were found for the first time several writings in an ancient language : Hittite. Its deciphering was completed by 1935, and it appeared that :

1) it was clearly linked to other old Indo-European languages like Latin, Greek and Sanskrit ;
2) in words that could be retraced to a PIE origin, it contained specific sounds at the very places where de Saussure had needed to postulate their existence.

This is a very strong result in a Popperian way of thinking [5] : while it isn't too difficult to invent a law that takes all observed phenomena into account, any hypothesis that can predict some phenomena that are still to be observed will be vindicated in the strongest possible way when they are indeed observed.

From then on, both the law of regular changes and the existence of PIE laryngeal sounds (whatever they were) were considered as certainties by a vast majority of linguists.

2. A look at comparative methodology

2.1. "Give me one hundred parameters

… and I'll build you an elephant. With the 101st , it will waive its trunk"[1]. This well-known epistemological aphorism means that if you create a law that is intricate enough, you'll always be able to describe the whole set of observed phenomena. A sketchy law will let some of them aside ; to the other extreme, a law which distinguishes each individual case of a finite list as different will obviously be "correct" in its description, but of no interest ; at some intermediate level of

[1] The unconvinced reader may try replacing "parameter" with "gene"

complexity, there will exist one law or set of laws which will describe all observed phenomena in the most parsimonious way.
We may replace "parameters" by "prime implicants" and see that the aim of QCA is the search of this "least intricate complete law".

Hard-core Popperianists, however, will see a flaw in this methodology : there is no way to check whether the laws you came at are a faithful description of reality. Especially as it is quite possible, as QCA-users know it, to get several laws which all describe the observed phenomena with (near-)equal parsimony.
They will argue that you need a supply of unstudied cases, to allow you to check your hard-obtained laws against and see whether they still hold, and to help you distinguish between several possible laws if needed.
In all honesty, we're siding with them.

2.2. A methodological suggestion

The tradition, when using the comparative method, is to use at once the whole corpus[2] at one's disposal and look, with intuition as one's only guide, for a law that may explain the form of all words (in QCA terminology, all observed cases).
One typical problem, and not too complicated, is the study of *splits*. We speak of a "split" when one sound from a former language seems to have evolved into two or more sounds in a more recent language that derives from it, according to its context in the word. This is the most plausible explanation when to one sound of one language correspond several sounds in another language spoken at the same time. We recognise the basic technique of the reconstructive method.
For example, OD [uo] (modern Dutch [u:][3],, spelled "oe") corresponds to OHG [o:] or [e:]. If this is the mark of a split, it means that there was a sound in Proto-Germanic, referred to, with all due care, as *[o:] which has evolved into [uo:] in OD, but has split, according to the context, into OHG [o:] and [e:].[4]
Looking at the corpus, one may remark that the OHG sound is [e:] when, and only when, the following vowel was [i], long or short (context of *i*-umlaut). By the way, the law is the same in OE, but the result in *i*-umlaut context is [oe:] rather than [e:].
Our observation yielded a simple law, which explains all cases taken into account.
Now, let's reason *a la* Popper, and check this law against some other cases. Oops … there are none.
Comparativists have always worked on the whole corpus at their disposal, and this, popperianists would say, is wrong. Barring the unexpected discovery of another set of cognate triplets for OE, OD and OHG (which is always possible, by looking at ancient texts), we don't have any way of falsifying (the key word in Popper's theory) our would-be-law, thus no way of establishing it on firmer ground, which would be possible had it resisted several attempts at falsifying it [5, p.36].
We'd like to suggest trying another approach : selecting some part of the corpus (covering all possible values of contextual variables, configurations being as far apart from each other as possible, as is usual in QCA applications), finding a candidate for the title of "law", and checking it against the rest of the corpus.

[2] The set of all known words of the language.

[3] A colon marks a long vowel.

[4] The problem is compounded by the fact that OG [e:] may also correspond to OD [ie:], but this is not relevant here.

Of course, it may happen that the law be falsified, and in this case we only need to work on another subset of the corpus (not necessarily disconnected from the one we used on our first attempt) and have another try at a solid law. But our approach to this second study will unavoidably be influenced by our partial results from the first one, and this is a bad thing. The solution to that problem is to let some machine -which, of course, won't let itself be influenced- do it for us. This is where QCA comes in.

3. QCA as a tool for comparative linguistics : an example

3.1. The study of splits using QCA

To avoid, as far as possible, the reader being influenced by one's knowledge of one's mother tongue, as would be quite possible if I took examples from a Romance language, English or German, I studied an esoteric case, a split that appeared in the transition from Proto-Germanic (**PG**) to Old Icelandic (**OI**)[5]. When studying a remote language (in time or space), the difficulties are increased, because the corpus we know of with sufficient certainty might be rather meagre. With Old Icelandic, however, this isn't a problem, because it's the language of a rich literature.

OI has two main types of infinitive forms : with or without inserted [j]. For example, we have *kjo:sa* "to choose" (by the way, those two words are cognates) and *wekja* "to awaken" (cognates once more). In PG, as it has been reconstructed, some infinitives also have inserted *[j], but they don't correspond to OI cases; For example, OI *her∂a* "to harden" corresponds to a PG infinitive that had inserted [j], as can be seen from its cognates, e.g. Old Saxon *herdian.*

We might therefore be observing the result of a split of some PG verbal ending – which could have been *[jan] or *[jana]- into two types of OI infinitive, one (type *wekja*) with inserted [j], the other (type *her∂a*)[6] being indistinguishable from forms (type *kjo:sa*) coming from PG infinitives without [j]. Or it might be something more intricate.

What we need is a way to determine which conditions from PG determine the presence or absence of PI [j]. This is a classical task for QCA. We strongly suspect that the presence of PG [j] is a necessary condition to obtain PI [j], but let's avoid unnecessary assumptions ; QCA should be able to tell us. What can the other conditions be ?

3.2. Choosing variables

We'd rather be honest : specialists of Germanic languages have studied this problem long ago, and came with an answer :

OI inserted [j] arises whenever PG had inserted [j] AND one of the following conditions is fulfilled :

(i) the [j] came after a sequence "short vowel - single consonant" ;

*(ii) the [j] came after a guttural consonant, i.e. *[k] or *[γ]*[7].

Since the sound preceding [j] always was a consonant, (*i*) may be rephrased as :

The penultimate sound before [j], was a short vowel.

[5] Information about OI is taken from [8d].

[6] PI ∂ is similar to English *th* in "this".

[7] *[γ] is believed to have been similar to Dutch *g*.

Our intent being to check whether the answer obtained through the use of "traditional" methods is right, it is obvious that the identity of the penultimate sound, the presence of PG [j] and the articulation point of the consonant should be taken into account. It is equally obvious that, in order to avoid the objection of having only worked with conditions that we know will be useful, we'll need to throw a few red herrings into the study. These will be chosen among conditions that are known to influence the apparition or perpetuation of some sound in other languages.
In the general case, however, the possible conditions will be far too numerous to be taken into account simultaneously, and even if some version of QCA allowed us to do it, it would be a bad idea, because it would create far too much logical cases [1, p.39].

The selection of "interesting" conditions may only be guided by knowledge of similar problems from the same language or another, or general knowledge of possible linguistic changes. Of course, QCA allows us to start with a large number of possible conditions and use only a few of them in each of a series of runs.

Here, however, we found it reasonable to limit ourselves to the following conditions :

- presence of PG [j] ;
- articulation of the preceding consonant : labial, dental or guttural. As this variable has 3 possible values, it needs to be dichotomised ;
- nasalisation of the preceding consonant ;
- type of the penultimate sound : long vowel, short vowel or consonant. Also needs to be dichotomised ;
- the fact that the action described by the verb needs active participation from the subject (as is illustrated by the difference between "look" and "see").

We then got 7 variables : Jpg, Dent, Gutt, Nas, Cons, Short, Act.
If comparatists were right, the reduced expression for "1" outcomes, i.e. the presence of [j], should be JPG * (GUTT + SHORT), which would appear as (JPG*GUTT) + (JPG*SHORT)[8].

3.3. Specificities of the truth tables

With 7 variables, one would expect to have 128 (2^7) possible combinations. However, remember that two pairs of variables (Dent/Gutt and Cons/Short) are the mark of a dichotomization of variables that originally had 3 possible values. Both variables in one pair may never be true at the same time. This means that there are in a way two kinds of "logical cases" :

- those which could have been observed, but weren't [1, p.60] ;
- those which can't be observed because they contain a contradiction, as would be the case for CONS*SHORT.

This reduces the number of possible cases to 72 (2^3*3^2). However, the "impossible" cases can be treated as classical logical cases by QCA. The coming of software that treat variables with more than two values, like TOSMANA, will allow us to reduce our variables to five.

[8] With the convention, used in [1], of writing a condition in uppercase letters when the formula uses it with value "1".

But the essential point is this one : if one takes as granted the law of regular changes – and we will-, there can be *no* contradictory row.
When there is one, it means that the same set of conditions could have a positive or negative outcome. This goes directly against the fundamental law.
Nowadays, nearly every linguist would reason as de Saussure did, and react to the presence of contradictory outcomes by hypothesising that one omitted some important factor from the set of conditions rather than putting the regularity paradigm in doubt.

3.4. Solving contradictions

When a contradictory row appears in the study of a problem from Political or Social Science, there may be more than one explanation :

(i) one ore more relevant variables were omitted from the study ;
(ii) the problem isn't deterministic, that is, the same conditions may indeed produce different outcomes ;
(iii) the threshold value used for dichotomising a numeric variable was wrongly chosen.

In the case of the study of phonetic changes, only the first explanation is possible.
This makes it the perfect example of a principle that, consciously or not, underlies case studies under QCA : the primary aim of such a study is to check whether the set of variables taken into account is complete ; and when faced with a contradictory row, one should at first assume it wasn't, and look for other possible explanatory variables [1, pp.78-79]. Here, it is the only possible reaction.

3.5. Adapting the corpus

Another check that can, and should, be made is to vary the part of the corpus submitted to analysis and check whether the reduced expression obtained at the end of QCA processing is the same each time. This is a good attempt at falsifying the first result ; if one's formula passes the test, if it remains the same with another set of examples, it becomes rather believable.

3.6. When more than one reduced expression appears

It often happens that several equally parsimonious reduced expressions are produced by QCA processing. In this case, the usual reflex is to look at their meaning and choose, among them, the one that makes the most sense to the researcher, based on one's own knowledge of the problem under study.
In the case of linguistic changes, this is also true. However, one should first perform new QCA processings, with other parts of the corpus. It could well happen that one study gives formulae A and B, while another gives B and C. In this case, B "must" be the right answer, as it is the only one compatible with all the data. If, however, there still remain more than one possible formula, then it is time to use likelihood as a tool for deciding the "right" formula [1, p.79].
One important consideration is that, in many languages, similar sounds undergo similar changes in similar contexts. For example, in Corsican, the conditions for the "softening" of unvoiced stops into voiced stops on one hand, and of voiced stops into continuous consonants on the other, are the same (neither after a pause, nor after a stressed syllable, nor after a consonant) [8b]. Stating that similar changes will appear

in similar contexts is too drastic, but when one of the formulae obtained through QCA is the same as one obtained in the study of another, similar, change in the same language, its likelihood is greater.

3.7. Finding new hypotheses for phonetic laws

We performed the study described in *3.2*, using 18 examples from [8d][9], both with and without the inclusion of logical cases.

Without logical cases, the formula was too complicated to be of any use, as might well be guessed from the fact that the cases under study had been chosen in such a way as to be as far from each other as possible.

Allowing QCA to use logical cases, a rather parsimonious law was found. But we were on for a surprise : while the problem under study had received a solution in the hands of grammarians :

JPG * (GUTT + SHORT) or (JPG*GUTT) + (JPG*SHORT),

the formula we got was slightly different :

GUTT + (JPG*SHORT).

Not being ready to assume that our study or the grammarians' must have overlooked something essential, we asked ourselves whether those two formulae could be equivalent.

It is apparent that, if GUTT happens only when JPG does, they are, because, in this case, JPG * GUTT is the same as GUTT only.

Logicians will agree with the statement that :

$(G \Rightarrow J) \Rightarrow [(J \wedge G) \Leftrightarrow G]$.

We then looked at the list of 18 examples and found that there were only three cases with GUTT ; al three of them had also JPG. A look at an extended corpus confirmed the implication GUTT ⇒ JPG, that is, when a PG infinitive's root ended with a guttural, its complete form contained inserted [j]. This is quite plausible : the evolution (at an earlier stage) of the ending [-kana], if it ever existed, into [-kjana], makes sense for articulatory reasons. Similar changes have been observed in many languages.

This is QCA at its best. The comparison between the "expected" response and the result of QCA processing made apparent an unsuspected (at least to us) law about the set of variables : GUTT doesn't appear without JPG. Perhaps this had been noticed before, through other means, but here it seems a very natural product of the QCA specific methodology of re-organising information [1, pp.79-80].

QCA appears, once again, as a wonderful tool for helping the researcher satisfy Albert Szent-Györgyi's definition of scientific discovery : "seeing what everybody has seen and thinking what nobody has thought".

BIBLIOGRAPHY

[1] G. DE MEUR & B. RIHOUX, *L'analyse quali-quantitative comparée*, Bruylant-Academia, Louvain-la-Neuve 2002.

[2] A. FOX, *Linguistic Reconstruction*, Oxford Un. Press 1995.

[9] 18 examples are only a small part of all existing cases (several hundred verbs), but they were initially chosen in [8d] to cover the largest possible range of values for context variables ; for this reason, we thought them to be numerous enough.

[3] A. MEILLET, *Introduction à l'étude comparative des langues indo-européennes*, 1934.
[4] C. PEETERS, *Les indo-européanistes connaissent-ils la méthode comparative ?*, in *Modèles linguistiques et idéologies : Indo-Européen II",* S. VANSEVEREN Ed., pp. 81-86, Ousia, 2002.
[5] K.R. POPPER, *Conjectures and Refutations, the Growth of Scientific Knowledge*, Routledge, London 1963.
[6] C.C. RAGIN, *The Comparative Method : moving beyond Qualitative and Quantitative Strategies*, Un.California Press 1987.
[7] F. de SAUSSURE, *Mémoire sur le système primitif des voyelles dans les langues indo-européennes,* Vieweg, Paris 1879-1887.

[8] Student's notes from the following courses given at the Université Libre de Bruxelles :
[8a] M. DOMINICY, *Linguistique I : phonétique, phonologie, morphologie*, 1995-1996.
[8b] M. DOMINICY, *Grammaire comparée des langues romanes,* 1996-1997.
[8c] F. MAWET, *Grammaire comparée des langues indo-européennes I,* 1997-1998 & *II*, 1998-1999.
[8d] C. PEETERS, *Grammaire comparée des langues germaniques*, 1999-2000.