Cladistic methods in linguistics and Dollo’s Law

Lukas Reinarz
Supervisor: prof. dr. Helen de Hoop
Radboud Universiteit Nijmegen
Contact: l.reinarz@web.de

ABSTRACT

Cladistic methods used for making phylogenetic family trees of biological species are an important tool in evolutionary biology and linguistics. They are based on the assumption that a group of species sharing the same genetic features (genotypes) must have evolved from the same common ancestor and that such features cannot come back once vanished. However, language change can be cyclic and a law in evolutionary biology, Dollo’s law, states that only features that are not genetically coded (phenotypes) can evolve in a cyclic way. Since linguistic features are phenotypic, cladistic methods used in linguistics are not reliable.

Keywords
Cladistics, phylogenetics, language evolution, linguistic family relationships, Dollo’s law, genotypes, phenotypes.

INTRODUCTION

Language change and linguistic family relationships are a highly studied area of linguistics. By trying to find features shared by different languages, linguists want to be able to tell how closely related language are. Some languages are close members of a language family, such as German and Dutch, others are not related to each other, for instance Finnish and Quechua.

In order to find out how closely related languages are, one uses methods with which one can create phylogenetic family trees, called cladograms. These methods originally come from evolutionary biology which uses them to investigate which species genetically belong together and which species share a common ancestor.

This is also done for languages. Dunn, Greenhill, Levinson and Gray (2011) present cladograms of four large language families (Austronesian, Indo-European, Uto-Atzekian and Bantu) which are based on word order. They say that by “[d]rawing on the powerful methods developed in evolutionary biology, we can […] track correlated changes during the historical biology of language evolution as languages split and diversify” (Dunn et. al., 2011, p.79). Thus, they say that biological methods are suitable for languages.

However, there are linguistic phenomena that do not occur in biological evolution, such as cycles. Van Gelderen (2009) notes that one can speak of a linguistic cycle when a linguistic feature such as a word is replaced by another one and when this pattern repeats itself over time. So, when a word keeps being replaced by another word, one calls this a linguistic cycle. In biology, *Dollo’s law* (1893) states that only features which are not coded in the genes of a species can change cyclically. So, genetically coded features cannot.

The issue now is that in evolutionary biology, one works with genetic features, *genotypes*, and these do not change in a cyclic way, according to Dollo’s Law. In language, however, features that can change cyclically are used for cladograms, but these cannot be genotypic but have to be *phenotypes* as they can be influenced by external factors.

In this paper, I will elaborate on the hypothesis that phenotypic features of language cannot be used as data for cladograms because they are influenced by the (linguistic and social) environment.

CLADOGRAMS IN LINGUISTICS

Languages are subject to changes. Therefore, it is possible to establish family relationships between languages by using the method of *phylogeny*. According to Brinkman and Leipe (2001, p. 323), in phylogeny one studies evolutionary relations. This method originally comes from biology and has the aim of classifying species in terms of families and to identify how closely related language are to each other. Dunn (2014, p. 190) says that phylogenetic methods “can be applied to any domain which varies according to general evolutionary processes”.

A direct link between biology and linguistics is the identification of *homologous features* in languages (Platnick & Cameron, 1977). Features are called homologous when they share the same origin. In biology, these are gene sequences coming from the same common ancestor (Brinkman & Leipe, 2001). When two languages share some homologous features, they are related to each other since they share the same features from a common ancestor. Because of this, it is possible to make cladograms from such features.

In order to find such homologous features, one has to examine whether two features are “similarities of the entire system” (Platnick & Cameron, 1977, p. 383) meaning that the features really have to be attributed to a common ancestor and not only look the same due to coincidence. An example of homology from the Romanic languages is the similarity of the numerals *un*, *une* in French, *uno*, *una* in Italian and *uno*, *una* in Spanish.

These words are a homologous feature of the three languages which is supported by the fact that the words for *two* and *three* are similar to each other as well (Fr. *deux, trois,* Sp. *dos, tres,* It. *due, tre*). Figure 1 shows a possible clustering of French, Italian and Spanish based on the numerals for *one*, *two* and *three*. It is clear that the Italian and Spanish numerals show a great similarity with each other and that French is more different. So, Italian and Spanish have to be clustered together whereas French has its own branch in the cladogram. Because the numerals of all three languages are very similar to each other, they are assumed to have a common ancestor.

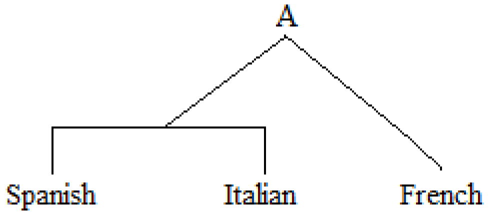

Figure 1: A cladogram with French, Italian and Spanish. This is a possible cladogram made of the numeral for "one", “two” and “three” in the three languags. The letter "A" represents the protolanguage, i.e. the common ancestor of the languages.

Homologous features in biology are situated in the genes of species and so, genetics can tell us how closely related some species are. The question now is whether it is justified to use biological methods in order to identify homologous features in language.

CLADOGRAMS IN BIOLOGY

Phylogenetic methods in biology

Genetic features shared by some species are called *homologous*. But in order to use cladistics methods for establishing family relations between species, one has to examine if a feature is *primitive* or *derived.* Derived features are features which can be seen in two or more species but which were not present in their common ancestor (Brinkman & Leipe, 2001). Species which share a derived feature are related to each other but not to other species in which that feature is still in its primitive state which means that the feature was already present in the common ancestor.

An example of this is the following. Horses have one digit and apes and kangaroos have five. They all are mammals. However, lizards have five digits, too. So, lizards, horses and kangaroos share a homologous feature, the number of digits. The only possibility to explain for this is that all these species have one common ancestor. In the course of evolution, the number of digits of horses has changed.

Once the derived and primitive features have been identified, one can make a cladogram. One method for doing so is the *Henning-method* (Lipscomb, 1998). Let us assume that there are three species, called A, B and C. Let there are four features be subject to analysis. Every species is assigned the values 0 (primitive) or 1 (derived) for every feature. Table 1 shows the distribution of the features in the species. The term *outgroup* refers to the species with the same common ancestor which are not classified and do not share any of the features.

	F1	**F2**	**F3**	**F4**
Outgroup	0	0	0	0
A	1	0	0	0
B	1	1	0	1
C	1	0	1	1

Table 1: The distribution of the features in the species. 0 means that a feature is primitive, 1 means that it is derived.

What can be seen from this table is that all species share one feature (F1) which is not present in the outgroup which suggests that they belong together. The first step is to split the outgroup from species A, B and C. Then, F2 is derived in species B, so, B has to be split off the rest. The same is true for species C and feature F3. F4 is derived in species B and C, so the splitting point has to be between A and the two species. The resulting cladogram is shown in Figure 2. The numbers indicate the number of the step taken.

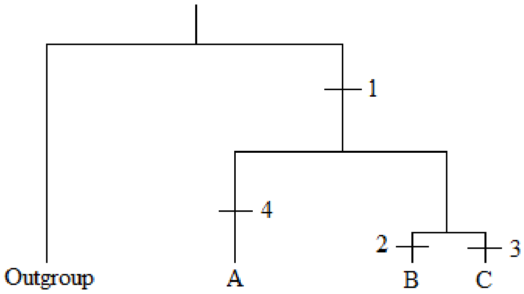

Figure 2: The cladogram made from the data in Table 1.

An important point is that the features that are used have to be in the genes of the species classified. Genes show the evolutionary history of a species and are therefore suitable for identifying family relations between species. Those features are called *genotypes* and are altered by genetic mutations alone. Features which change under the influence of external factors such as the environment are called *phenotypes* and are not encoded in the genes.

With respect to the difference between genotypes and phenotypes, Dollo (1893) formulated a law, which has become known as *Dollo’s Law*. I will come to this law in the following section.

Dollo’s Law

Dollo’s Law states that features that have vanished in the course of evolution will not appear again. Evolution is irreversible according to Dollo’s Law. This means that a feature that a species has lost will never reoccur, not even if the species lives in exactly the same environment as when the feature was still present.

Dollo notes that functional and physiological features in fact do reoccur but structural and morphological features do not (Gould, 1970). Furthermore, Dollo’s Law is only applicable to genotypes. Phenotypical features can reoccur because they are not dependent on genes but on the environment in which a species lives. An example of this is human body size. When there is little food, human beings are shorter than when there is enough food. So, when the situation changes, this can influence body size.

The explanation for Dollo's Law may be that the possibility that a lost feature reoccurs is extremely small. The genetic changes that took place in the course of millions of years would have to change precisely in the opposite order in order to get the lost feature back. This possibility is approximately zero. Irreversibility thus is a matter of chance.

Dollo's Law is important when talking about phylogenetic methods in linguistics. In the following section, I will show why this is the case. It will turn out that, taking into account Dollo's Law, the use of biological methods for linguistic purposes is not justified.

LANGUAGE AS AN ORGANISM

Linguistic cycles

Since Dollo's Law tells us something about the absence of cyclical changes in the evolution of species, it is interesting to examine how that law relates to language change. Languages do not change linearly but in a cyclic way. Van Gelderen (2009, p.9) defines a linguistic cycle as "a name for changes where a phrase or word gradually disappears and is replaced by a new linguistic item".

An example of such a linguistic cycle is the case cycle described by van Gelderen (to appear). Case is a morphological category (i.e. a category with respect to word form) which affects flexion of nouns in a language. In the phrase *John's car*, *John's* is a genitive case because it is marked by the genitive ending (affix) -*s* meaning that John possesses the car. In the course of the history of a language with cases, one can observe frequently that case affixes vanish. The function the affixes carry in the clause (*possessor* in the case of *John's car*) cannot be conveyed anymore so that other means are required to express the same function. This happens by means of prepositions, so that *John's car* becomes *the car of John*.

So, case endings can be replaced by prepositions. These prepositions gradually become grammatical items which are placed behind the noun and are now called *postpositions*. Then, these postpositions are more and more attached to nouns and cannot be separated from them anymore. They have become case affixes again. In Figure 3, this cyclic change is illustrated.

An example of a language in which a case ending is replaced by a preposition is English, as in *John's car* (which has a genitive ending) and *the car of John*. In Turkish, there is a postposition, *ile* meaning '(together) with', which is placed behind a noun and which has been becoming a case ending. Speakers of Turkish can, for instance, say *Mehmet-le* which means 'together with Mehmet'. The fact that the sound *i* vanishes from the postposition makes it even more plausible that the postposition has been grammaticalizing.

That language change can be cyclic has an important consequence for using cladistics methods in linguistics. This consequence is explained in the following section.

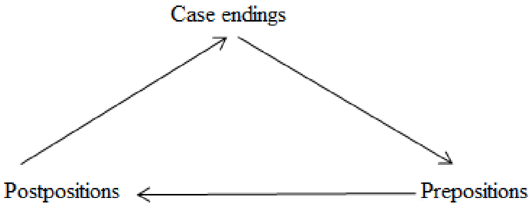

Figure 3: A schematic overview of the case cycle. When case endings vanish, they are replaced by prepositions which are placed behind the noun in the course of the process. These postpositions again become case endings.

Linguistic features as phenotypes

As outlined above, Dollo's Law is only applicable to phenotypes. This is important when talking about language change because, since language changes cyclically, the linguistic features that change cannot be genotypic but have to be phenotypes, as follows from Dollo's Law. This is logical because there are several linguistic phenomena that influence language change but that never appear in biological evolution.

One of those phenomena is language contact. Language contact is frequently mentioned as one of the most important motors for language change (Thomason, 2001; Lee, 1987; Silva-Corvalán, 1994; Cabral, 2003; Heine & Kuteva, 2005; Miestamo, Sinnemäki & Karlsson, 2008). Aikhenvald (2002) describes an example of contact-induced language change in a language called Rituarã from the Yucuna language family. This language has case affixes to express locative case (places) which is not typical of Yucuna languages. According to Aikhenvald, this change took place due to language contact with the Arawak languages which show locative case regularly.

Another factor which can influence language change is culture. In communities with compulsory school attendance and relatively large differences in education of the population, the effect of prescriptive grammar is often very large and language changes are often stigmatized (Drake, 1977).

These factors show that language change is to a large extent subject to external factors. From a biological point of view, this is only possible when the features that change are phenotypic. This and Dollo's Law indicate that linguistic features are not genotypic, but phenotypic. The consequence of this is explained in the following section.

Back to the beginning: Cladograms in linguistics

We have seen that linguistic features that are used for comparing languages are phenotypic but that the methods used for this are based on genotypic features. As a consequence, those linguistic features cannot be used for cladograms. The conclusion is that there are no reliable cladograms of languages because the data that is used is not reliable. The precise form of a cladogram is dependent on the choice of the data. So, using cladistics methods in linguistics is based on the wrong assumption that the features used are genotypes.

This also holds for the cladograms made by Dunn et al. (2011). The authors themselves show that linguistic features they use (the order of adposition-noun and verb-object) do not change unidirectionally but that change can be cyclic. So, these features are phenotypes and their cladograms are not reliable.

CONCLUSION

We have seen that using cladistic methods from biology is not justified in modeling linguistic interrelationships. The reason for this is that the features that are used for those analyses are subject to external changes and factors such as culture and language contact. What is more is that the features change in a cyclic way which is only possible for phenotypic features, according to Dollo's Law.

Cladograms such as in Figure 1 are thus no reliable representation of linguistic family relations. The numerals *one*, *two* and *three* that are used for the analysis have their specific shapes because of the environment in which the three languages changed. It is not the case that the 'genes' of the languages have changed.

Although cladistic methods are more and more used in linguistics, it has turned out that the assumption on which the use of these methods is based is not reliable. In order to examine how reliable representations of linguistic relations can be made, further research has to be done that takes into account that features that are subject to change are phenotypic rather than genotypic.

ROLE OF THE STUDENT

Lukas Reinarz was an undergraduate student of Linguistics supervised by Prof. Dr. Helen de Hoop while carrying out the research outlined in this paper. The topic was worked out by the student and the supervisor. Lukas read all the relevant literature and informed Prof. Dr. de Hoop about the content. The ideas and conclusions that are described in his thesis and outlined in this paper were mostly initiated by him. The writing was done by Lukas as well.

ACKNOWLEDGMENTS

I would like to thank my supervisor Prof. Dr. Helen de Hoop who supported me with her very valuable comments and ideas.

REFERENCES

1. Aikhenvald, A. Y. (2002). *Language contact in Amazonia*. Oxford: Oxford University Press.
2. Brinkman, F. S. & Leipe, D. D. (2001). Phylogenetic analysis. *Bioinformatics: a practical guide to the analysis of genes and proteins* , 323-358.
3. Cabral, A. S. A. C. (2003). *Contact-induced language change in the western Amazon: The non-genetic origin of the Kokama language* (Academic dissertation). UMI, Ann Arbor.
4. Dollo, L. (1893). The laws of evolution. *Bulletin de la Société Belge de Géologie, de Paléontologie et d'Hydrologie, 7*, 164-166.
5. Drake, G. F. (1977). Evolutionary linguistics. *Annual Review of Anthropology, 37*(1), 219-234.
6. Dunn, M. (2014). Language phylogenies. In C. Bowern & B. Evans (red.), *The Routledge handbook of historical linguistics* (pp. 190-211). New York: Routledge.
7. Dunn, M., Greenhill, S. J., Levinson, S. C. & Gray, R. D. (2011). Evolved structure of language shows lineage specific trends in word order universals. *Nature*, *473*(7345), 79-82.
8. Gelderen, E. van (2009). *Cyclical Change*. Amsterdam: John Benjamins Publishing.
9. Gelderen, E. van (to appear). The dependent marking cycles: Case. In Elly van Gelderen (ed.), *Linguistic cycles*.
10. Gould, S. J. (1970). Dollo on Dollo's law: irreversibility and the status of evolutionary laws. *Journal of the History of Biology, 3*(2), 198-212.
11. Heine, B. & Kuteva, T. (2005). *Language contact and grammatical change.* Cambridge: Cambridge University Press.
12. Jespersen, O. (1917). Negation in English and other languages. *Høst & Søn.*
13. Lee, J. R. (1987). *Tiwi today: a study of language change in a contact situation*. Canberra: Dept. of Linguistics, Research School of Pacific Studies, The Australian National University.
14. Lipscomb, D. (1998). Basics of cladistic analysis. *George Washington University*.
15. Miestamo, M., Sinnemäki, K. & Karlsson, F. (2008). *Language complexity: Typology, contact, change*. Amsterdam: John Benjamins Publishing.
16. Platnick, N. I. & Cameron, H. D. (1977). Cladistic methods in textual, linguistic, and phylogenetic analysis. *Systematic Biology, 26*(4), 380-385.
17. Silva-Corvalán, C. (1994). *Language Contact and Change: Spanish in Los Angeles*. Oxford: Oxford University Press.
18. Thomason, S. G. (2001). *Language contact*. Edinburgh: Edinburgh University Press.