Cultural evolutionary modeling of patterns in language change : exercises in evolutionary linguistics

Landsbergen, F.

Chapter 6

Reconstructing the diachrony of *krijgen* with synchronic data using phylogenetic inferencing techniques

6.1 Introduction[1]

Phylogenetics is the study of historical relationships between species in biology. These relationships are reconstructed by looking at similarities and differences between the species in the present, and by assuming that, through time, new species develop out of existing ones by a process of mutation and selection. The similarities and differences between species can be studied at two levels: the morphological and the molecular level. The former deals with any relevant formal, phenotypic characteristics, such as warm- or cold-bloodedness, the shape of the beak, or the presence of a backbone. The latter involves sequences of DNA or RNA.

Outside biology, phylogenetic inferencing techniques have recently become increasingly popular in comparative linguistics to study the historical relations between languages within larger language families, such as Indo-European (Warnow 1997, Gray & Atkinson 2003, McMahon & McMahon 2003, Bryant, Filimon & Gray 2005, Nakleh, Ringe & Warnow 2005), Bantu (Holden, Meade & Pagel 2005), Austronesian (Gray & Jordan 2000, Greenhill & Gray 2005) and Papuan (Dunn, Terrill, Reesink, Foley & Levinson 2005).

These studies are often met with suspicion by 'conventional' linguists. As McMahon & McMahon (2005: 26) put it, there is often the fear that 'historical linguists, with their deep knowledge of individual languages and groupings [are replaced] by sleek, humming computers and programs which smooth out all the

[1] I would like to thank Michael Dunn, who has introduced the phylogenetic method to me during the 2007 LOT Winterschool in Nijmegen, and who has given me valuable advice on the analysis.

bumps'. However, computational methods should not be considered to be a replacement, but instead to be an additional tool in historical research that can be used for testing hypotheses or gaining insight. For example, the phylogenetic study by Gray & Atkinson (2003) was carried out in order to find quantitative support for one of two conflicting hypotheses about the origin of Indo-European, the 'Anatolian farming hypothesis' and the 'Kurgan expansion' theory. Similarly, the study by Gray & Jordan (2000) supported one of several hypotheses about the colonization of the Pacific by Austronesian-speaking people. Also, the computer programs that carry out the analyses have to be fed with manually collected material, which usually consists of lists of cognates. The choice whether or not two words are considered cognates is not done by the computer, but has to be done by a trained linguist. Therefore, linguistic expertise will remain needed at all times, and computers will not take over the field.

Going back to biology, phylogenetic studies can be carried out on very diverse scales. On the largest scale, there are the relationships between the major lineages of life, such as eukaryotes and bacteria. On a smaller scale, there are the relationships between the kingdoms such as plants, animals and fungi, and on an even smaller scale, there are the relationships between for example groups of vertebrates such as fish, dinosaurs and mammals. Even within one species, phylogenetic methods can be performed, for example to reconstruct the origin and divergence of the human race (Cavalli-Sforza, Menozzi & Piazza 1994).

In linguistics, phylogenetic studies are usually carried out on languages that are part of a larger language family like Indo-European or Austronesian (see references above), but, similar to biology, there is no a priori reason to restrict oneself to this particular scale. An example of a study on a slightly smaller scale is Minett & Wang (2003), who look at the relationship between seven dialects of Chinese.

The possible use on different scales leads to the question whether phylogenetic methods can also be used to study the development of one particular linguistic item within a single language, as is the typical object of study in historical linguistics and grammaticalization research. As an example, let us consider the English word *while*. This word today exists as a noun, a temporal marker and a concessive marker. These three uses co-exist today, but this has not always been the case. The noun *while* is the original use of the word, and from this noun, a temporal marker developed (while the original noun remained in use). In a later phase, a concessive marker developed from the temporal marker. This means that, analogous to the cases for species and languages described above, we have an item showing descent with modification, leading to a situation of variation in the synchronic state. With the use of the phylogenetic method, it might therefore be possible to

reconstruct the historical relationship between these items. In other words, one could reconstruct a word's history on the basis of its current synchronic variation.

In this chapter, I will explore the usability of this approach. The linguistic item I will use for this exploration is the Dutch word *krijgen*. In this chapter 4, I have discussed both the synchronic variation of the verb and its historical development and I will use this as a basis for the enterprise in this chapter. I will focus on two main questions: (1) is it possible to get a reliable reconstruction of the development of *krijgen* on the basis of synchronic data, and (2), what are the limitations of this approach?

In the next section, I will first give a more detailed description of the phylogenetic method in general. In section 6.3, I will define the different uses of *krijgen* in terms of this method, followed by results in 6.4 and a discussion in 6.5.

6.2 Phylogenetic inferencing methods

In this section, I will give a brief introduction to phylogenetic inferencing methods in general and their use in linguistics in particular. Although these methods are gradually becoming more popular in linguistics, the explanatory literature on the topic is still almost solely written for a biological audience (e.g. Felsenstein 2004, Ridley 2004). An introduction into the method that is specifically meant for linguists is McMahon & McMahon (2005), and in several other linguistic studies the main principles are briefly introduced as well (e.g. the collection of articles in Mace, Holden & Shennan 2005). Still, linguists that are interested in the subject will need to resort to the biological literature as well.

Phylogenetics is the study of historical relations between items, such as species or languages. These historical relations can be studied in different ways and hence there are different phylogenetic techniques that are used. In this respect, two main approaches can be distinguished: so-called 'character-based' and 'distance-based' methods. Character-based methods use characters such as morphological features or DNA-sequences from the relevant species.[2] Each of these species can be defined by the state of a series of these characters (e.g. 'present' or 'absent' for morphological features and C, T, G and A for DNA). A tree is then reconstructed in which, in the smallest number of steps, the character sets of all species can be fitted, thus giving insight in the most likely order in which these species are related historically.

Distance-based methods also use these character sets, but instead reduce these sets into single values representing the distance between one species and the

[2] Note that in phylogenetic research, the term 'characters' is used in the sense of 'characteristics'.

other. A tree can then be reconstructed that accounts for the actual evolutionary distance between species, or languages.

To illustrate both methods with an example, let us consider four randomly chosen species: a frog, a rat, a chimpanzee and a human. For each species, or taxon, a set of (relevant) characteristics can be gathered. We want to reconstruct the history of these four taxa on the basis of this data: how are the four species related?

Distance-based methods

Table 1 is a made-up example of a set of characters, and shows a sequence of 12 bases of mitochondrial DNA for our four species, in which each base can be either A, C, G or T.

	1	2	3	4	5	6	7	8	9	10	11	12
human	A	A	C	T	G	G	A	G	T	C	A	G
chimpanzee	A	A	C	T	A	G	G	G	C	C	A	G
rat	G	A	T	G	A	G	T	G	A	C	T	G
frog	A	C	T	G	A	C	T	G	A	C	T	G

Table 1. A fictitious representation of a DNA sequence of four species.

In the distance-based method, the relationships between the species are calculated by looking at the distance between the taxa. The distance between two taxa is the number of steps (or mutations) it takes to get from one taxon the other. This information is shown in the distance matrix (table 2).

	human	chimpanzee	rat	frog
human	-			
chimpanzee	3	-		
rat	7	6	-	
frog	8	7	3	-

Table 2. Distance matrix of the four species. The values indicate the number of mutations to get from the DNA sequence of one species to another.

We can now construct a tree on the basis of the values from the distance matrix (figure 1). In this tree, the branch lengths represent the distances between the taxa.

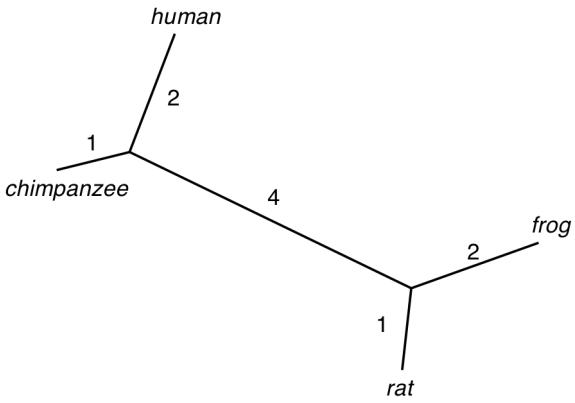

Figure 1. An unrooted phylogenetic tree based on the distances of the four taxa. The values and branch lengths indicate the distances between the taxa from the distance matrix in table 2.

The tree in figure 1 is a so-called 'unrooted' tree: it only provides us with information about the distance between the taxa and does not give any information about the order of the taxa in a tree. The usual way to add the latter is to use an 'outgroup root': a taxon that is known to predate the taxa that are subject of study. Figure 2 shows the rooted version of the tree from figure 2, using 'drosophila' as the outgroup root.

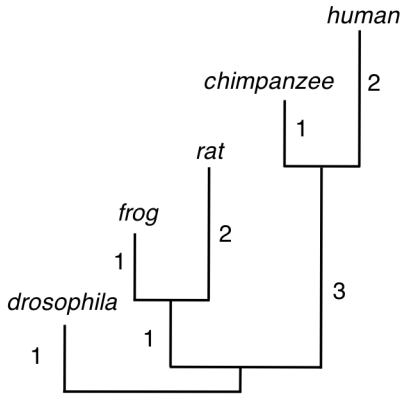

Figure 2. Rooted tree with 'drosophila' as the outgroup root.

There are different algorithms to construct trees from a distance matrix. The most commonly used are UPGMA and neighbor-joining. The neighbor-joining algorithm is an iterative algorithm which clusters the two closest taxa under one node, and then

goes on to calculate other distances, treating the two clustered taxa as one single taxon. UPGMA stands for Unweighted Pair Group Method using Arithmetic Mean. Like neighbor-joining, it also uses an iterative clustering algorithm, but the methods differ in the exact way of clustering and the calculation of distances. Another major difference with neighbor-joining is that UPGMA is based on the assumption that evolution within all branches of the tree takes place at the same rate. This is known as the molecular clock hypothesis in biology (Zuckerkandl & Pauling 1962). Although I explained before that distance-based methods produce unrooted trees, this is not true for the UPGMA-method. In this algorithm, the constant rate assumption makes it possible to reconstruct the historical order of the branching (needless to say: only if the assumption about the rate of evolution is valid).

A special kind of distance method is the NeighborNet method (Bryant & Moulton 2002). This method does not produce trees but networks instead. Normally, distance methods will always create a tree, even if support for branches might be weak. The NeighborNet method will not necessarily give one optimal tree as output, but allows for a network structure in which several alternative trees are suggested.

Character-based methods

Another way to reconstruct the relationship between taxa is the character-based method. Unlike distance-based methods, this method does not reduce the character sequence (e.g. as the one shown in table 1) to a single distance value. Instead it aims to produce a tree that can account for all character sequences in the smallest number of steps.

As an example, let us consider our four species frog, rat, chimpanzee and human again. For this method, the fictitious DNA sequences from the previous examples can again be used. However, other types of characters can also be used for both methods, and to illustrate this I will use morphological and behavioral characters in this example. These characters are usually presented with binary coding, giving either absence or presence of each character. Table 3 gives an example.

	tool use?	cold blooded?	vertebrae?	language?
human	1	0	1	1
chimpanzee	1	0	1	0
rat	0	0	1	0
frog	0	1	1	0

Table 3. An example of binary coded morphological and behavioral characters for the four species.

In trying to infer the relationships between the four example species, we reconstruct the steps that lead from the character sequence of one taxon to the other, using the information from table 3. A step would be either 0 > 1 or 1 > 0: it takes one step to get from human to chimpanzee, two steps to get from human to rat, one step to get from rat to frog, etc. Example 1 shows the steps to get from the human character sequence to the sequence of the other species:

1)	a.	human	1 0 1 1	>	1 0 1 0		chimpanzee					
				1								
	b.	human	1 0 1 1	>	0 0 1 1	>	0 0 1 0		rat			
				1		2						
	c.	human	1 0 1 1	>	0 0 1 1	>	0 0 1 0	>	0 1 1 0	frog		
				1		2		3				

With this information, the unrooted tree in figure 3 can be reconstructed, in which the intermediate steps are represented by vertical lines.

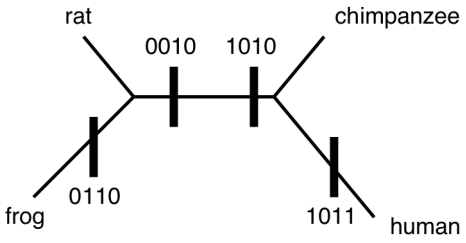

Figure 3. An unrooted phylogenetic tree of four species based on a set of four characters. The vertical lines indicate transitions between the sets.

As is the case for the distance-based methods, the trees that can be reconstructed in character-based methods are basically unrooted and an outgroup root can be added to construct a rooted tree. Using the taxon ‘drosophila’ again (with the fictitious character sequence 0100) as our outgroup root, this would lead to the rooted tree shown in figure 4.

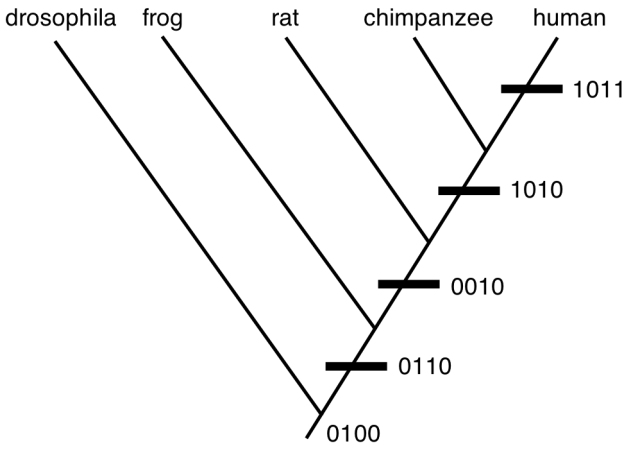

Figure 4. Rooted phylogenetic tree based of five species based on a set of four characters. The taxon 'drosophila' is added as an outgroup root.

In this example, it is obvious which single tree can be reconstructed from the data. However, with more taxa and more characters, this reconstruction becomes more complicated, as more trees become possible. As for distance-based methods, several different character-based methods have been developed, such as the method of maximum parsimony and that of maximum likelihood. For maximum parsimony, trees are given a 'score' and this score is determined by the number of steps required in the tree. The tree with the least number of steps is said to be the most parsimonious tree. The method of maximum likelihood resembles the method of maximum parsimony in that it also assign scores to a set of possible trees, yet it differs from the former method in that it does not necessarily aims at finding the 'shortest' tree (i.e. the number of steps), but at finding the tree with the highest statistical probability.

A special kind of character-based method that has been increasingly used in recent years, also in linguistic research (e.g. Gray & Atkinson 2003, Holden, Meade & Pagel 2005) is *Bayesian inference*. Bayesian methods are described by Holden et al. (ibid.: 60) as 'not [a] search for the best tree(s) according to an optimality criterion, but instead [as sampling] a large number of trees in proportion to their likelihood. Trees with a higher likelihood will be sampled proportionately more often, but trees with an intermediate likelihood which occur more frequently in the universe of possible trees, will also be represented in the sample. The aim is to represent phylogenetic uncertainty in the sample.'

The method is usually used together with a so-called Markov Chain Monte Carlo (MCMC) simulation. Basically, MCMC stands for the actual searching of a sample of the tree space, "preferably steering toward those trees which maximize a value called the 'posterior probability'" (Wichmann & Saunders 2007: 390). This posterior probability reflects the probability that a given tree is the correct one for the dataset. With the continuing search in the tree space, a point will be reached when this posterior probability will not improve. A sample of trees from beyond this point is then collected, from which a consensus tree is constructed. Each branch in the consensus tree has a score to represent its strength, representing the number of times the node is found in the tree sample. Branches for which there is too little support (those that appear in the tree sample under a certain threshold level, which is usually set at 50 percent), are left out of the consensus tree.

Phylogenetic methods in linguistics

In the study of historical relationships between languages, both distance-based methods (e.g. Swadesh 1950, Gray & Jordan 2000 for Austronesian) and character-based methods (e.g. Ringe, Warnow & Taylor 2002 and Gray & Atkinson 2003 for Indo-European) have been used.

In these studies, the groups that are usually compared to one another are languages from a single language family such as Indo-European, and the character sets used in these studies are usually based on lexical data, counting the number of cognate words in the different languages. Table 4a gives an example of such a cognate list for the words of four semantic categories in five randomly chosen Indo-European languages (based on table 5.1 in Bryant, Filimon & Gray 2005: 70). In table 4b, the same information is represented in binary form, with languages that share a specific cognate receiving the value 1, and languages that do not the value 0.

	'father'	'foot'	'four'	'fish'
Greek	pateras	podhi	teseris	PSARI
Irish	athair	COS	cesthair	iasc
Spanish	padre	pie, piede	quattro	pesce
Riksmål	far	BEN	fire	fisk
Dutch	vader	voet	vier	vis

Table 4a. Cognate list of four semantic categories in five Indo-European languages. Non-cognates are shown in capital letters.

	'father'	'foot'	'four'	'fish'
Greek	1	1	1	0
Irish	1	0	1	1
Spanish	1	1	1	1
Riksmål	1	0	1	1
Dutch	1	1	1	1

Table 4b. The data from table 4a in binary form.

The choice whether or not two languages share the same cognate has to be made manually, and as table 4a suggests, a good knowledge of the comparative method is necessary. Also, one has to keep in mind the possibility that words can be borrowed from another language without the two languages necessarily having to be related (see Minett & Wang 2003 and Warnow, Evans, Ringe & Nakleh 2004 for methods of taking borrowing into account).

Phylogenetic methods in this study

In this chapter, I will consider three different phylogenetic methods on a data set of the Dutch verb *krijgen* and discuss how the results of the methods are compatible with the findings from chapter 4. These methods are the distance-based methods NeighborNet and neighbor-joining, and a Bayesian analysis, which is used in character-based methods. This choice is partially inspired by a study by Wichmann & Saunders (2007), who tested several methods on a dataset of Native-American Languages and concluded that these three methods gave the most reliable results.

The three analyses are carried out using two open source software packages: the Splitstree program[3] (Huson & Bryant 2006) for the NeighborNet and neighbor-joining analyses, and the BayesPhylogenies package (Pagel & Meade 2004)[4] for the Bayesian analysis.

[3] http://www.splitstree.org/
[4] http://www.evolution.rdg.ac.uk/BayesPhy.html. Extensive information about the different packages that are on the web can be found on the PHYLIP-website by Joseph Felsenstein: http://evolution.genetics.washington.edu/phylip.html

6.3 Defining taxa and characters for *krijgen*

The problem of defining taxa and characters for a single linguistic item

As I explained in the previous section, constructing a phylogenetic tree is a two-step process. First, each taxon is characterized by a set of characters and second, the relationships between taxa are calculated on the basis of this character set.

In the reconstruction of historical relationships between linguistic items within a single language, the first step in the process already leads to a problem. Contrary to the phylogenetic reconstruction of species within a larger family of species, or of languages within a language family, the taxa in the case of different uses of a linguistic item are not known beforehand. How should one proceed?

Of course, in defining taxa and characters in the case of a single linguistic item such as *krijgen*, we are not starting completely from scratch: we can use both the information embedded in the synchronic variation of the item and rely on general knowledge about the development of comparable items. For example, for a modal verb such as *can*, it is known that these kinds of verbs often change in the kinds of subjects (animate vs. inanimate) and complements (noun phrases vs. verb phrases) they take. Therefore, it is useful to define taxa on the basis of these properties. If it turns out that some characterizations do not lead to any useful distinctions, they can be left out at a later stage. Also, one could use the knowledge that in modal verbs such as *can*, different uses such as main verb, mental ability, ability and root possibility constructions are distinguished. This knowledge can be combined with information about the synchronic variation of the item.

Another complexity is the choice of an outgroup root if we want to produce rooted trees. For *krijgen*, the intransitive use is an obvious candidate for an outgroup root. However, this means we need the independently motivated knowledge that the intransitive use indeed predates the transitive. For the case of *krijgen*, this knowledge can be found in most historical dictionaries. It can be argued that using this knowledge does not interfere with the goal of the reconstruction itself, which is to gain insight in how the different transitive uses of the verb are historically related. That is, we would accept the generally known historical order of development (intransitive > transitive) and use the phylogenetic reconstruction to focus on the development of the latter part of the development. Strictly speaking, however, using an outgroup necessarily means relying on more than just synchronic data.

In the next section, I will define the characters and taxa of *krijgen*, using the knowledge of its synchronic variation that I discussed in the chapter 4.

Taxa and characters for krijgen

When focusing first on the transitive use of *krijgen*, the following 'constructions' can be identified on the basis of the direct object:

2)	*Concrete object-construction:*	een cadeau krijgen (van iemand) 'to get a present (from someone)'
	Abstract object-construction:	antwoord krijgen (van iemand) 'to get an answer (from someone)'
	Internal objects-construction:	griep krijgen 'to get the flu'

The construction with abstract objects can be further refined when looking at the role of transfer. In (3), *antwoord* 'answer' necessarily involves (metaphoric) transfer from one person to another, but in (4), no transfer takes place and *krijgen* only seems to have aspectual meaning: a 'change into a state of having'.

3) De hoogleraar kreeg een onduidelijk antwoord van de student.
'The professor got an unclear answer from the student.'

4) Ik hoop dat we dit jaar op vakantie beter weer krijgen.
'I hope we will get better weather on our holiday this year.'

Within the construction with concrete objects, a subdivision is also possible when focusing on the role of the subject, which has a recipient role in (5) and an agent role in (6).

5) De zoon kreeg voor zijn verjaardag van zijn vader een auto.
'The son got a car from his father for this birthday.'

6) De criminelen deden alles om de auto te krijgen.
'The criminals did everything to get the car.'

Finally, in the construction with internal objects, a subdivision is possible between those objects that can be possibly be controlled by the subject (7), and those that cannot (8).

7) De vrouw kreeg na jaren van yogalessen eindelijk de gemoedsrust die ze al zo lang had gezocht.
'After years of yoga classes, the woman finally got the peace of mind that she had been looking for for so long.'

8) Elk voorjaar kreeg de jongen flinke hooikoorts.
'Each spring, the boy got bad hay fever.'

Apart from the transitive constructions, *krijgen* is also used in combination with various adjuncts, like in the resultative construction (9), the aspectual use of *krijgen* with a participle (10-11, with a difference in the grammatical subject of the second verb), the semi-passive (12) and the use with *te* + infinitive (13).

9) Met dit nieuwe afwasmiddel krijgt men zelfs de meest vieze pannen weer schoon.
'With this new detergent, one gets even the dirtiest pans clean again.'

10) Lukt het je nog dit mailtje vandaag verstuurd te krijgen?
'Will you still be able to get this mail sent today?'

11) De gijzelnemer kreeg zijn eisen ingewilligd.
'The hostage taker got all his demands complied with.'

12) De volkszanger kreeg de koninklijke onderscheiding uitgereikt door de burgemeester.
'The popular singer got his royal decoration handed out by the mayor.'

13) De deelnemers kregen pas na lange tijd de uitslag te horen.
'The contestants were only told the result after a long time (litt. 'got to hear the result').'

These eleven different constructions (shown in examples 3-13) can be considered as the taxa that are needed for the reconstruction. They now need to be further defined, using a set of characters.

In general, the more characters that can be used in the description of the different taxa, the better, because it will lead to a more refined characterization of each taxon. For example, the word list used in the study of Indo-European by Gray & Atkinson (2003) consisted of 200 items, and the list of structural features used in the study of Papuan languages by Dunn et al. (2005) consisted of 125 items. Of course, only those characters should be added that are relevant in distinguishing the taxa (which obviously means leaving out those characters that render similar values for all taxa). Furthermore, it is also important to minimize correlations between characters, although it is impossible to rule out all covariance.

For the characterization of the *krijgen* taxa, it will not be possible to obtain a set of characters in the order of magnitude of those mentioned above. The question is whether this is a major obstacle in the pursuit of getting a reliable reconstruction of the historical development of the verb *krijgen*. It could very well be possible that

some linguistic items are more suitable for this particular line of research than others, since they allow for a higher number of characters that can describe them.

Starting with characters that relate to the subject, the following questions can be listed:

14) I Is the subject animate?
II Is the subject agentive?
III Can the action be the intention of the subject?
IV Is there a change of state of the subject?
V Is the subject of krijgen the subject of the main verb?

These questions give yes (1) and no (0) answers. Questions (I-II) are straightforward. Question III is necessary to distinguish uses like *geluk krijgen* 'to get luck' from uses like *koorts krijgen* 'to get fever'. In both cases, the subject is not the agent, but objects like *geluk* 'luck' can be obtained by an intentional subject, while objects like *koorts* 'fever' cannot.

Uses like *een waarschuwing/antwoord krijgen* 'to get a warning/an answer' mostly go with animate subjects because of the nature of the direct object. This is not the case for the use with abstract objects like *goed weer krijgen* in which there is no transfer, and for the use with concrete objects, as examples (15-16) show.

15) De voorlichtingscampagne kreeg een feestelijke start.
'The information campaign got a festive start.'

16) Bij het inchecken krijgt de bagage een nieuw label.
'At check-in, the baggage gets a new label.'

A change of state of the subject, as in question IV, occurs in uses with an internal object, like *koorts/geluk krijgen* 'to get a fever/luck'. Question V is specifically relevant for aspectual use of *krijgen*. In (17), the main verb of the sentence is *versturen* 'to send', and the subject of this verb is also the subject of *krijgen*. This is not the case in (18), in which *inwilligen* 'to comply with' is not done by the hostage taker, but by another, implicit participant.

17) (=10) Lukt het je nog dit mailtje vandaag verstuurd te krijgen?
'Will you still be able to get this mail sent today?'

18) (=11)De gijzelnemer kreeg zijn eisen ingewilligd.
'The hostage taker got all his demands complied with.'

The list of characters can be extended with a set of questions that relate to the used direct objects.

19) VI Is the direct object a concrete object?
VII Does the direct object become the possession of the subject?
VIII Is the object controllable?
IX Is there a change of state of the direct object?

These questions also give yes (1) or no (0) answers, and all are straightforward. The resultative uses and the semi-passive mostly get a concrete direct object (e.g. *de piano de trap op krijgen* 'get the piano up the stairs'), while this is not the case for the use with the *te* + infinitive (e.g. *het antwoord te horen krijgen* 'get to hear the answer'). The group of controllable objects largely overlaps with that of concrete ones, with the exception of objects like *geluk* 'luck' and *genade* 'mercy'. Direct objects undergo a change of state in the resultative use, in which the changed state is expressed by an adjunct phrase.

A final set of questions relates to different aspects, such as the presence or absence of an adjunct phrase (as in the resultative use) or another participant (as the entity who is responsible for the actual transfer). Question XIII distinguishes the resultative use with PP or AP adjuncts from the aspectual use with verbal adjuncts.

20) X Is there a compulsory adjunct phrase?
XI Is an extra participant needed for the transfer?
XII Is there (metaphoric) transfer?
XIII Does *krijgen* describe the main action?

In order to produce a rooted tree, it is necessary to add an outgroup root to the list of taxa. As I explained in the previous section, the intransitive use of *krijgen* (which has the meaning 'to fight, to strive for') is a proper candidate. It is distinguished from the other uses by adding a final question to the list:

21) XIV Is there a direct object?

Together, this leads to the following table of taxa and characters (table 5 on the next page). I will apply this dataset to three different phylogenetic methods, and compare the outcomes with the results of the diachronic study of *krijgen* from the previous chapter.

		I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV
1	INTRANSITIVE *Ø krijgen* 'to make an effort'	1	1	1	0	1	0	0	0	0	0	0	0	1	0
2	CONCR_OBJ_1 *het zwaard krijgen* 'to get the sword'	1	1	1	0	1	1	0	1	0	0	0	1	1	1
3	CONCR_OBJ_2 *een cadeau krijgen* 'to receive a gift'	0	0	0	0	1	1	1	1	0	0	1	1	1	1
4	ABSTR_OBJ_1 *goed weer krijgen* 'to get good weather'	0	0	0	0	1	0	0	0	0	0	0	0	1	1
5	ABSTR_OBJ_2 *antwoord krijgen* 'to get an answer'	0	0	1	0	1	0	0	1	0	0	1	1	1	1
6	INT_OBJ_1 *geluk krijgen* 'to get luck'	0	1	1	1	1	0	0	1	0	0	0	1	1	1
7	INT_OBJ_2 *griep krijgen* 'to get the flu'	0	0	1	1	1	0	0	0	0	0	0	0	1	1
8	RESULTATIVE *het slot open krijgen* 'to get the lock open'	1	1	1	0	1	1	0	1	1	1	0	0	1	1
9	ASPECTUAL_1 *eisen ingewilligd krijgen* 'to get the demands complied with'	1	1	1	0	0	1	0	1	1	1	0	0	0	1
10	ASPECTUAL_2 *het mailtje verstuurd krijgen* 'to get the mail sent'	1	1	1	0	1	1	0	1	1	1	0	0	0	1
11	SEMI-PASSIVE *de prijs uigereikt krijgen* 'to get the prize handed out'	0	0	0	0	0	1	1	1	0	1	1	1	0	1
12	TE+INFINITIVE *een geheim te horen krijgen* 'to get to hear a secret'	0	0	1	0	1	1	0	0	0	1	1	1	0	1

Table 5. The taxa of *krijgen* with their corresponding character values. A description of each character is found in the text above. Each taxon has been given a short name that will be used in the actual phylogenetic study.

6.4 Results

In this section, I will discuss the results of three methods of phylogenetic reconstruction, NeighborNet, neighbor-joining and Bayesian analysis and compare these results to the development of *krijgen* in chapter 4. I will start this section with a short overview of the findings of this chapter.

A short overview of the development of krijgen

In the previous chapter, I have discussed the history of *krijgen* based on diachronic data. *Krijgen* started as an intransitive verb 'to fight, to make an effort', with an agentive subject. From this use, the first two branches are a transitive variant with the meaning 'to obtain by effort' and a resultative variant with the meaning 'to get X at/from location Y', both appearing before 1400. The transitive use is initially combined with concrete objects like 'sword' and 'chess board'. Later, this set of objects is extended to mental and physical states of the subject and abstract objects (*wille* 'consent', *antwoord* 'answer').

Over time, the transitive use loses its subject agentivity, and this process starts with the use with 'state' objects, followed by the use with abstract objects and, lastly, with concrete objects.

From the resultative use, new uses also develop: first there is an extension from locative complements to PP-complements describing states. Later, these 'state' complements also allow the use of participles, which can be seen as the first aspectual use of *krijgen*. All these uses keep an agentive subject, which runs contrary to the development of the transitive use.

There are two auxiliary uses of *krijgen*, and both appear relatively late in the verb's history. The *te* + infinitive construction is first found in the 18th century and is used with a non-agentive subject. This use seems to have developed in a process of specification from the transitive use with a non-agentive subject.

The second auxiliary use is the semi-passive, and this use is the latest development of *krijgen*, appearing first in the early 20th century. I hypothesized that this use developed from the aspectual use of *krijgen*, in an environment in which the prototypical (transitive) use of the verb had become almost completely non-agentive.

Next, I will subject the dataset from the previous section to three different methods of analysis, and compare the outcomes with the diachrony of *krijgen* mentioned above

NeighborNet

NeighborNet is a distance-method that produces networks instead of trees. This means that it does not produce one optimal tree, but a network in which several alternative trees are suggested as well.

Figure 5 shows the resulting network, in which the names of the different taxa are those found in table 5. Obviously, the network does not have a very strong tree-like structure, which is due to the fact that most taxa can be linked to each other in different ways. This means that, on the basis of the given data, there is no strong support for a single possible relationship between the different uses of *krijgen*.

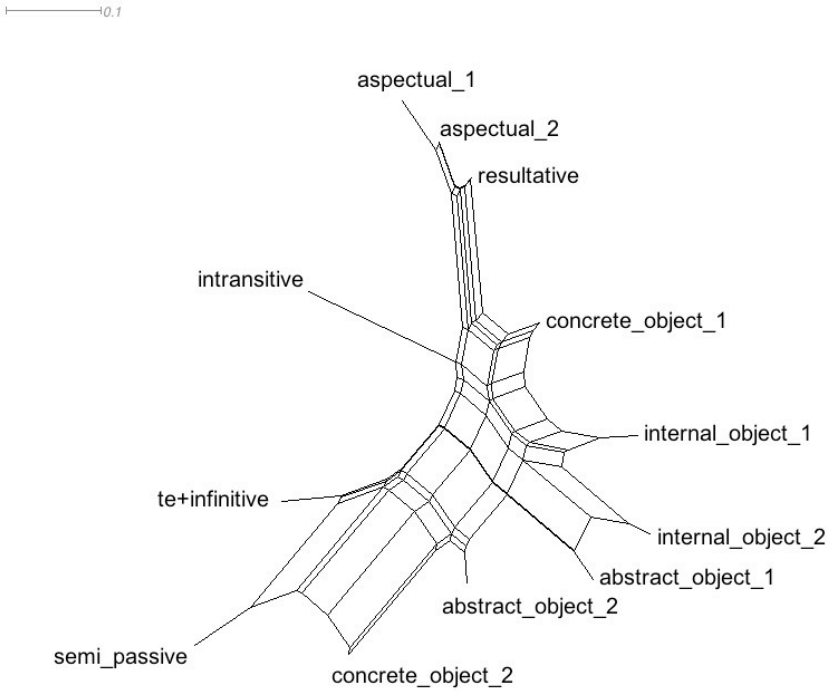

Figure 5. NeighborNet analysis of the 12 taxa of *krijgen*. The branch lengths indicate relative distance.

The most strongly supported 'branch' is that of the resultative and aspectual uses. The network correctly suggests that the aspectual use has developed from the resultative. The order of the two aspectual uses is also as was hypothesized in the

previous chapter: first, the subject of the participle is similar to the subject of *krijgen* (aspectual_2), while later the two subjects can be different (aspectual_1).

The network does not provide any strong clues about the relationships between the early transitive uses of *krijgen* in the middle of the graph. Still, based on the relative distances between these different uses in the graph and the fact that the intransitive use is the original use of *krijgen*, there is some (yet weak) support for the path of development that was also found in the diachronic corpus study:

22) concrete objects with an agentive subject (concrete_object_1)
|
internal objects with an agentive subject (internal_object_1)
|
internal objects with a non-agentive subject (internal_object_2)
|
abstract objects with a non-agentive subject (abstract_object_1)
|
abstract objects with a recipient subject (abstract_object_2)
|
concrete objects with a recipient subject (concrete_object_2)

Also, the network does show some support for a grouping of uses with an agentive subject (in the top part) versus those with a recipient subject (in the bottom part), with non-agentive use in the middle. This intermediate position of the non-agentive use gives some support for the hypothesis that *krijgen* first developed its non-agentive use from its agentive use, and that the 'receive' sense developed from this non-agentive use later, as was found in the corpus study.

Interestingly, the semi-passive and *te* + infinitive use are grouped with the 'receive' senses in the bottom left part of the network. This supports the hypothesis that both uses have developed from the 'receiving' use.

Neighbor-joining

The neighbor-joining analysis is one of the methods that construct a tree on the basis of the relative distances between the different taxa. It clusters the two closest taxa under one node, and then goes on to calculate other distances, treating the two clustered taxa as one single taxon. Strictly speaking, the resulting neighbor-joining tree therefore does not provide any information about the history of *krijgen*. However, the distances between the different constructions can be used as an indication of their historical relatedness, using the assumption that the closer two constructions are in distance, the closer related they are in time. These distances are

represented in the branch lengths in the tree. Neighbor-joining trees can be both unrooted and rooted when an outgroup root is assigned.

Contrary to the NeighborNet network, neighbor-joining analysis always produces a tree, even when there is little support for certain branches. A way to obtain an idea how well the branches in the tree are actually supported by the data is to perform a so-called bootstrapping procedure (Holder & Lewis 2003: 279, Felsenstein 2004: 334ff).

Bootstrapping randomly resamples the data set to produce pseudo-replicate data sets, and the tree-building algorithm is then repeated on these replicate sets. Pseudo-replicate data sets have the same data as the original data set, but the order of the data has been altered. By also searching for trees in these 'extra' sets, it can be measured how many times each branch in the original tree is recovered, thus giving a measure of its robustness. These values are shown in the trees in figures 6 and 7.

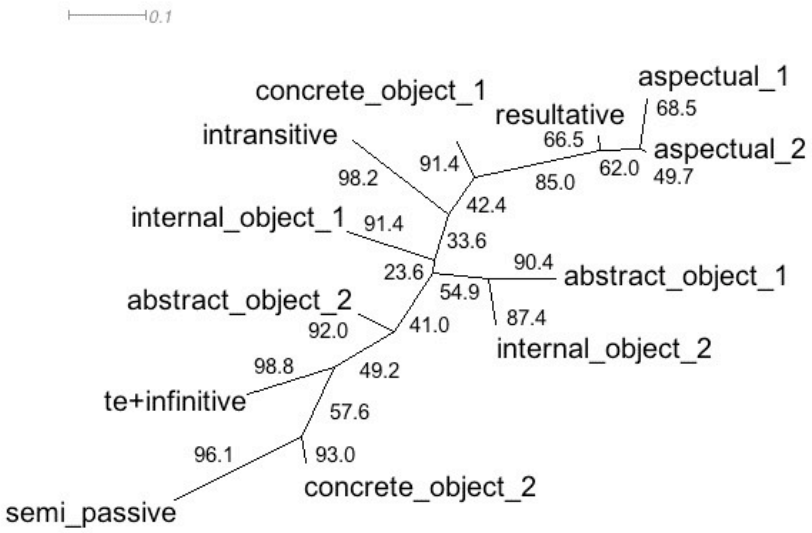

Figure 6. Unrooted consensus tree of 1000 bootstrap replicates, using the neighbor-joining method. The branch lengths indicate relative distances between taxa. The values on each branch show the percentage that the branch is present in the replicate set.

The unrooted tree (figure 6) strongly resembles the NeighborNet-network: the uses with an agentive subject are found close to the original intransitive use in the top part of the graph, the uses with a recipient subject in the bottom left part of the

graph, and the non-agentive use in the right middle. Again, the semi-passive and the *te* + INF uses are grouped with the recipient uses of *krijgen*.

Although many of the branches have considerably high bootstrap values, the branches that differentiate the main uses of *krijgen* (agentive subject / non-agentive subject / recipient subject) do not. Their bootstrap values never exceed 50, which means they are supported by the data in less than half of the cases. This finding is comparable to that of the NeighborNet analysis.

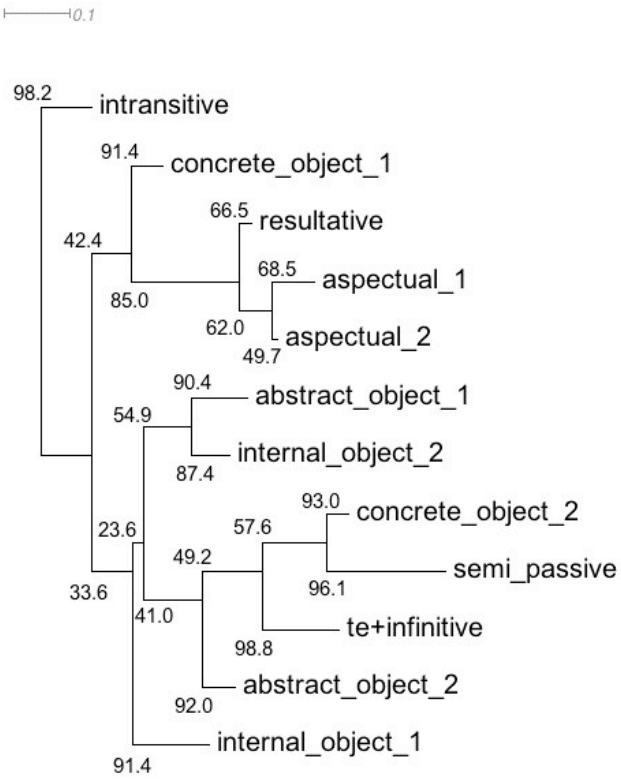

Figure 7. Rooted consensus tree of 1000 bootstrap replicates using the neighbor-joining method. The branch lengths indicate relative distances between taxa. The values on each branch show the percentage that the branch is present in the replicate set. The intransitive use is added as outgroup root.

Figure 7 shows the rooted version of the neighbor-joining tree, with the intransitive use as outgroup root. Basically, it holds the same information as the unrooted tree in figure 6, with the addition that the branches are now positioned based on their

relative distance to the root. The horizontal axis can therefore be taken as a representation of time, moving from old (left) to new (right). In general, the tree reflects the historical development reasonably well, with the use with an agentive subject (both transitive and resultative) appearing first, followed by the development of non-agentive use, and the use with a recipient subject after that. The semi-passive, which was found to be one the newest constructions in the diachronic corpus, also appears in the tree as the newest use, having developed from the use with concrete objects and a recipient subject ('concrete_object_2' in the graph).

However, when looking at the tree in more detail, some of the branches turn out to be problematic. The first split in the tree creates a branch of strongly agentive uses in the top part of the graph and a branch of less agentive uses in the bottom part of the graph. This split runs contrary to the hypothesis that the use with an internal object and an agentive subject ('internal_object_1' in the graph) has developed from the use with a concrete object and an agentive subject ('concrete_object_1'). The tree never shows a direct development of one use from another, like the aspectual use with a VP as adjunct developing out of the resultative use with a PP as adjunct, although such a development is possible in a phylogenetic tree. Instead, related uses share a common ancestor. Another example of this is the non-agentive use with an internal object ('internal_object_2'), that shares a common ancestor use with the original agentive use with an internal object ('internal_object_1').

Bayesian analysis

As I explained in section 6.2, Bayesian analysis samples a large number of trees in proportion to their likelihood, using a so-called Markov Chain Monte Carlo (MCMC) simulation. From this sample, a consensus tree is constructed, and each branch is this tree is given a score that represents the number of times the node is found in the tree sample.

The first step in the Bayesian analysis is therefore to search the space of possible trees. Figure 8 shows the results of this search, with the posterior probabilities of the sampled trees quickly reaching a point of convergence. It was found that 10,000 generations and a sampling of every 10^{th} tree was enough to reach a reasonable point of convergence (these values can differ strongly, depending on the number of taxa and characters in the dataset). The last 250 trees were kept, giving a so-called 'burn in' of 750 trees before the convergence point. From these 250 trees, a consensus tree was made using Splitstree.

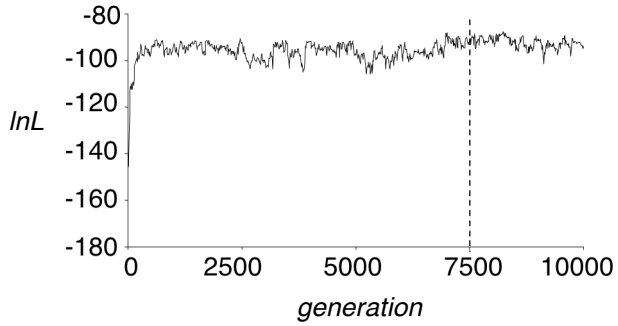

Figure 8. Posterior probabilities scores of 10,000 generations of trees, using a Markov Chain Monte Carlo simulation. The score for each 10th tree is shown. The dotted line indicates the point in which the score has reached a reasonable convergence.

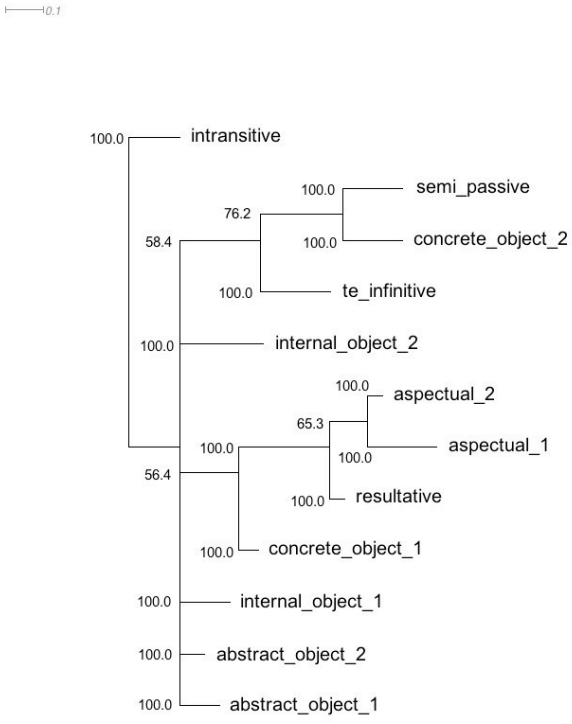

Figure 9. Majority-rule consensus tree based on the sample of 250 trees that was obtained with the MCMC simulation. The tree is rooted with the intransitive use. Branch lengths indicate distance, values are posterior probability scores.

The most obvious aspect of the resulting tree (figure 9) is that it has far less branching than the trees that were produced with the neighbor-joining analysis. There are only two major branches, one with the uses with a recipient subject in the top of the tree, and one with the uses with an agentive use in the middle, but even these two branches are poorly supported by the data. However, the branching within these main branches is similar to that in the neighbor-joining tree. For the rest, the Bayesian tree does not provide any information about the development of *krijgen*, or it has to be that no single path of development is supported by the data.

This result runs contrary to the finding of Wichmann & Saunders (2007: 391), who conclude that the results of the Bayesian analysis are superior to those of other methods. How can this difference be explained?

The major problem with the dataset I have used in this study is its size. Where cognate sets or typological features can give 100 or more characters, the number of characters in this study was only 12. This low number of characters obviously leads to branches that cannot be supported by a significant number of character values, as would be the case with a bigger character set. This is especially a problem for the Bayesian analysis, because this analysis works with a consensus tree that is constructed on the basis of multiple trees. With a low number of characters, the analysis will come up with differently structured trees that still have similar likelihoods, and these different structures will disappear in the consensus tree. Based on this study alone, it is therefore not possible to make a good judgment about the usefulness of the Bayesian method for this type of research.

6.5 Discussion

Phylogenetic inferencing techniques allow for historical research without historical data. It is therefore a possibly very useful tool in the study of linguistic change, both across and within languages. In this study, I have introduced the technique for the study of the development of linguistic items within a language. The first results are promising, yet the method has to be improved in order to become more useful.

The biggest challenge in using phylogenetic techniques in the study of single linguistic items such as *krijgen* is to establish an objective and workable character set. First, the character set that I used for *krijgen* was very small (12 characters). This has lead to branches that are sometimes based on only a single shared character value, and to the fact that conflicting branches can also be supported. With a higher number of characters, relationships between groups will be constructed more reliably. The question is whether a significant increase in the number of characters is possible for this type of historical research. I am inclined to say it is, but with the recognition that this limits the number of linguistic phenomena

that can be investigated. The best candidates are therefore those linguistic items that have undergone significant changes on all linguistic levels: syntax, semantics, phonology and pragmatics, and thus allow for a large character set. Also, these diachronic changes have to be traceable in the synchronic variation.

A second problem with the character set is the possible correlation between characters. Ideally, characters should 'behave' independently from each other, but for the characters used in this study, it cannot easily be determined whether this is the case. For example, I have used separate characters for 'subject agentivity' and 'subject animacy', while a degree of overlap, and thus correlation, between the two is rather likely.

The third problem is related to the first two and has to do with the fact that there is a degree of arbitrariness in the choice of characters. For each linguistic item to be studied, a new character set will have to be determined, and there seems no clear objective way to decide which characters have to be selected.

In conclusion, I would argue that using phylogenetic techniques in the study of single linguistic items is a promising enterprise in historical linguistics, while keeping in mind that there are several difficulties that have to be dealt with. However, with the limitations that I discussed above in mind, the method can still be considered useful in that it can be used as a supporting method in actual historical research.