A DISCOVERY SYSTEM FOR COMPONENTIAL ANALYSIS OF KINSHIP TERMINOLOGIES

Vladimir Pericliev* and Raúl E. Valdés-Pérez **

*Department of Mathematical Linguistics**
Institute of Mathematics and Computer Science, bl.8, 1113 Sofia – Bulgaria
peri@banmatpc.math.acad.bg

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

Abstract: The paper argues for the prospects of machine discovery in linguistics by describing a system, called KINSHIP, which automatically performs componential analysis of kinship terms. We briefly outline the system and illustrate it with the componential analyses of Seneca, Swedish and Polish. We show some facilities of KINSHIP that make it a convenient tool for the kinship analyst and discuss some implications of this work for machine discovery in linguistics.

Keywords: kinship, componential analysis, machine discovery in linguistics.

1. INTRODUCTION

The recent advances of cognitive and computer science have led to a significant gain in our understanding of the scientific discovery process. Continuing the tradition of 17th and 18th century scientists-philosophers like Descartes and F. Bacon, who were firmly convinced in the possibilities of a science of discovery, to date computer programs have been developed to simulate important historical discoveries (cf. esp. Langley et. al., 1987, where significant parts of 18th and 19th century Chemistry are re-discovered) or to help researchers in making really novel scientific discoveries in fields like mathematics (Graffiti), social sciences (Tetrad) or

ISBN: 0 08 043 438X

chemistry (Mechem). A snapshot of very recent work is contained in the Special Issue of the *Artificial Intelligence* journal (April, 1997) (for an overview of the history and methodology of the field, cf. e.g. Valdés-Pérez , 1996).

Contemporary linguistics, unfortunately, did not follow the course taken in these sciences. We have elsewhere (Pericliev, 1995) treated the reasons for this in some detail. Suffice it to say here that a major cause was the severe, and often unjust, criticism of the influential transformational grammarians (and esp. N. Chomsky) directed at so-called "discovery procedures" of American descriptivists, resulting in a totally negative connotation associated with discovery ever since.

In this paper we argue for the prospects of machine discovery in linguistics by describing a linguistic discovery program, KINSHIP, which automatically performs componential analysis of the meaning of kinship terms, as well as some implications of that research (cf. also Pericliev, 1995 implementing the familiar induction methods of J. S. Mill). The componential analysis of kinship vocabularies is indeed a less active area of study than it used to be some decades ago, but it still offers serious challenges to the analyst that can be more successfully met, we believe, with the KINSHIP program we describe. Importantly, the program models a creative linguistic discovery task, as generally conceded (cf. eg. "...kinship analysis have a mind-teasing quality of mathematical puzzles. The only cure for bafflement is to think hard and hope that the light will dawn!" (Leech (1974: 237)).

The paper is structured as follows. Section 2 is an overview of the the KINSHIP system. Sections 3, 4, and 5 give componential analysis of Seneca, Swedish, and Polish, respectively. Section 6 describes some facilities of the system which make it a convenient tool for the study of kinship systems, and Section 7 discusses some implication of our research for the prospects of machine discovery in linguistics.

2. THE *KINSHIP* PROGRAM

The KINSHIP program, written independently in both Common Lisp and Prolog, accepts as input data the kinterms with their attendant kintypes, stated in a traditional notation: F=father, M=mother, B=brother, S=sister, s=son, d=daughter, H=husband, W=wife. Besides these primary genealogical relationships, the system employs some auxiliary symbols to denote the sex of the speaker (m=male, f=female), the relative age of alter with respect to ego (e=elder, y=younger), etc. The system outputs the componential analysis of kinship terms as semantic features.

Here, for the sake of brevity, we shall only briefly mention the functions of KINSHIP (a detailed linguistically-oriented description is Pericliev and Valdés-Pérez (submitted), and an AI-oriented one, reporting also a further development of the system, is Valdés-Pérez and Pericliev, 1997). Endowed with a set of semantic features like «generation», «sex», «lineality» (currently numbering 25), the system produces the *most parsimonious componential analyses*, in the sense that it uses:

1. the least number of overall features sufficient to discriminate all kinterms, and
2. the least number of features for each kinterm meaning description.

The system is also capable of producing all alternative componential analyses, and, in the case of discrimination failure, can invent derived, features, by means of combining old features via logic operations like conjunction, disjunction, implication, etc.

3. EXAMPLE 1: SENECA

As our first illustration, we use the data on the Ameridian language of the Iroquois tribe Seneca, taken from the classical paper by Lounsbury, 1964; see also Leech, 1974. The corpus of Seneca consanguineal terms is given on Table 1. This is actually the input to the KINSHIP program. The output is the componential analysis presented on Table 2.

Table 1 Consanguineal Kinship Terms in Seneca (input)

1.	ha?nih	F FB FMSs FFBs FMBs FFSs FFFBss
2.	no?yeh	M MS MMSd MFBd MMBd MFSd MMMSdd
3.	hakhno?seh	MB MMSs MFBs MMBs MFSs MMMSds
4.	ake:hak	FS FMSd FFBd FMBd FFSd FFFBsd
5.	hahtsi?	Be MSse FBse MMSdse FFBsse MFBdse FMSsse MMBdse
6.	he?ke:?	By MSsy FBsy MMSdsy FFBssy MFBdsy FMSssy MMBdsy
7.	ahtsi?	Se MSde FBde MMSdde FFBsde MFBdde FMSsde MMBdde
8.	khe?ke:?	Sy MSdy FBdy MMSddy FFBsdy MFBddy FMSsdy MMBddy
9.	akya:?se:?	MBs FSs MMSss FFBds MFBss FMSds MMBss MBd FSd
10a.	he:awak	ms mBs mMSss mFBss mMBss mFSss mMMSdss
10b.	he:awak	fs fSs fMSds fFBds fMBds fFSds fMMSdds
11a.	khe:awak	md mBd mMSsd mFBsd mMBsd mFSsd mMMSdsd
11b.	khe:awak	fd fSd fMSdd fFBdd fMBdd fFSdd fMMSddd
12.	heye:wo:te?	mSs mMSds mFBds mMBds mFSds mMMSdds
13.	hehso?neh	fBs fMSss fFBss fMBss fFSss fMMSdss
14.	kheye:wo:te?	mSd mMSdd mFBdd mMBdd mFSdd mMMSddd
15.	khehso?neh	fBd fMSsd fFBsd fMBsd fFSsd fMMSdsd

The analysis uses the following features:

1. The binary feature +/- male describing the sex of the relative.

2. The multi-valued numerical feature generation (in this specific case having the following three values: > (senior generation than ego), < (junior generation than ego), = (equal generation with ego))

Table 2 Semantic Components of Seneca Kinship Terms Discovered by KINSHIP (output)

1.	ha?nih 'my father'	>generation	+parallel	+male	
2.	no?yeh 'my mother'	>generation	+parallel	-male	
3.	hakhno?seh 'my uncle'	>generation	-parallel	+male	
4.	ake:hak 'my aunt'	>generation	-parallel	-male	
5.	hahtsi? 'my elder brother'	=generation	+parallel	+male	+senior
6.	he?ke:? 'my younger brother'	=generation	+parallel	+male	-senior
7.	ahtsi? 'my elder sister'	=generation	+parallel	-male	+senior
8.	khe?ke:?, 'my younger sister'	=generation	+parallel	-male	-senior
9.	akya:?se:? 'my cousin'	=generation	-parallel		
10.	he:awak 'my son'	<generation	+parallel	+male	
11.	khe:awak 'my daughter'	<generation	+parallel	-male	
12.	heye:wo:te? 'my nephew'	<generation	-parallel	+male	+male-ego
13.	hehso?neh 'my nephew'	<generation	-parallel	+male	-male-ego
14.	kheye:wo:te? 'my niece'	<generation	-parallel	-male	+male-ego
15.	khehso?neh 'my niece'	<generation	-parallel	-male	-male-ego

3. The binary feature +/- senior showing age difference within one generation between alter and ego.

4. The binary feature +/- male-ego showing the sex of the speaker.

5. The binary feature +/- parallel showing the equivalence of sex between the two kin of the generation above ego or alter (which ever of those is junior); this feature reflects the anthropological demarcation between parallel/non-parallel cousins (for a popular description of this feature, cf. Sampson, 1974).

In the case of Seneca (cf. Table 2), five overall features suffice to discriminate all kinterms (there are no smaller sets of features, available to the program, that can accomplish the demarcation, cf our parsimony criterion 1); the meaning of each kinterm is represented with two, three or four features (our parsimony criterion 2 implies that any deletion of a feature would destroy the demarcation, and any addition will lead to redundancy). KINSHIP in this case re-discovered the componential analysis proposed by Lounsbury, 1964.

4. EXAMPLE 2: SWEDISH

Below we look at the Swedish consanguineal kinterms. The data reveals a remarkable morphological regularity in the formation of kinterms: from the set of nuclear family terms *fader, moder, broder, syster, son, dotter* (exactly coinciding with our atomic genealogical relationships F, M, B, S, s, and d), all the remaining terms are formed by combining the latter (just as the atomic genealogical relationships are juxtaposed to form more complex relationships).

Table 3 Swedish consanguineal terms

1. fader	F
2. moder	M
3. farfar	FF
4. morfar	MF
5. farmor	FM
6. mormor	MM
7. son	s
8. dotter	d
9. broder	B
10. syster	S
11. farbror	FB
12. morbror	MB
13. faster	FS
14. moster	MS
15. sonson	ss
16. dotters	ds
17. sdotter	sd
18. dotterdotter	dd
19. brors	Bs
20. systers	Ss
21. brordotter	Bd
22. systerdotter	Sd
23. kusin	FBs MBs FSs MSs FBd MBd FSd MSd

KINSHIP produced the componential analysis given on Table 4. The analysis employs the following features:

1. The binary feature +/- male, showing the sex of the relative.

2. The multi-valued numerical feature genealogical distance, expressing the number of consanguineal links between ego and alter.

3. The multi-valued numerical feature generation (self-explanatory).

4. The binary feature +/- male-1st-link, showing the sex of the first connecting relative (=link).

Table 4 Componential analysis of Swedish

1. fader	+male generation=1 geneal_distance=1
2. moder	-male generation=1 geneal_distance=1
3. farfar	+male-1st-link +male generation=2
4. morfar	-male-1st-link +male generation=2

ISBN: 0 08 043 438X

5.	farmor	+male-1st-link -male generation=2
6.	mormor	-male-1st-link -male generation=2
7.	son	+male generation=-1 geneal_distance=1
8.	dotter	-male generation=-1 geneal_distance=1
9.	broder	+male generation=0 geneal_distance=1
10.	syster	-male generation=0 geneal_distance=1
11.	farbror	+male-1st-link +male generation=1 geneal_distance=2
12.	morbror	-male-1st-link +male generation=1
13.	faster	+male-1st-link -male generation=1
14.	moster	-male-1st-link -male generation=1 geneal_distance=2
15.	sonson	+male-1st-link +male generation=-2
16.	dotterson	-male-1st-link +male generation=-2
17.	sondotter	+male-1st-link -male generation=-2
18.	dotterdotter	-male-1st-link -male generation=-2
19.	brorson	+male-1st-link +male generation=-1 geneal_distance=2
20.	systerson	-male-1st-link +male generation=-1
21.	brordotter	+male-1st-link -male generation=-1
22.	systerdotter	-male-1st-link -male generation=-1 geneal_distance=2
23.	kusin	geneal_distance=3

An interesting characteristic of this componential analysis is the use of the feature pertaining to the sex of the first connecting relative. Considering the need to discriminate between the kinterms for the kintypes

FF and MF

FM and MM

FB and MB

FS and MS

ss and ds

sd and dd

Bs and Ss

Bd and Sd

one may feel tempted to posit three distinct binary features, viz. patrilateral (for instance, FF is +patrilateral, and MF is -patrilateral), fratrilateral (for example, Bs is +fratrilateral, and Ss is -fratrilateral), and filial (e.g.. ss is +filial, and ds is-filial). Equivalently as regards simplicity (or, better, lack of simplicity), one can posit a single, but now non-binary, feature, call it "direction of kinship", having six different values for F, M, B, S, s and d. The second approach has in fact been employed for the Danish language whose kinship system closely resembles Swedish by Kuznecov, 1974, writing:

"This feature direction of kinship transmits information concerning the member of the kinship group through whom the kinship relationship signified by the given terms is realized: kinship on the father's side, on the mother's side, on the daughter's side etc." (p. 9).

Despite the fact that our system also disposes of the features patrilateral, fratrilateral and filia;, in order to ensure the use of a minimal number of overall features needed, it employs male-1st-link which is actually sufficient to make all the discriminations, which is by the way fairly obvious from the above listing of kintypes.

5. EXAMPLE 3: POLISH

The Polish data (cf. Kaluzin and Skorohodko, 1963) includes both the consanguineal and affinal kinterms.

Table 5 Polish consanguineal and affinal kinterms

1. ojciec	F
2. matka	M
3. syn	s
4. còrka	d
5. siostra	S
6. brat	B
7. siostra-stryjeczna	FBd
8. siostra-cioteczna	FSd MSd
9. siostra-wujeczna	MBd
10. brat-stryjeczna	FBs
11. brat-cioteczna	FSs MSs
12. brat-wujeczna	MBs
13. wuj	MB
14. stryj	FB
15. ciotka	MS FS
16. bratanek	Bs
17. siostrzeniec	Ss
18. bratanica	Bd
19. siostzenica	Sd
20. dziad	FF MF
21. babka	FM MM
22. pradziad	FFF FMF MFF MMF
23. prababka	FFM FMM MFM MMM
24. wnuk	ss ds
25. wnuczka	dd sd
26. prawnuk	sss sds dss dds
27. prawnuczka	ddd ssd sdd dsd
28. maz	H
29. zona	W
30. œwiekr	HF

ISBN: 0 08 043 438X

31. teœæ	WF
32. œwiekra	HM
33. teœciowa	WM
34. szwagier	HB WB
35. szwagrowa	HS

KINSHIP employs the six features male, generation, genealogical distance, male-1st-link, male-2nd-link, and affinal to accomplish the kinterms demarcation (cf. Table 6). The first four features are already familiar; the binary feature +/- affinal reflects whether the kintype is relative by marriage or by blood, and the binary feature +/- male-2nd-link shows whether the second link in the kintype is male or not (e.g. the kintype FFM is + male-2nd-link, whereas FMF is -male-2nd-link).

Table 6 Analysis of Polish with the alternative simplest kinterm definitions

1.ojciec	1. +male-1st-link generation=1 geneal_distance=1 -affinal 2. +male generation=1 geneal_distance=1 -affinal
2. matka	1.-male-1st-link generation=1 geneal_distance=1 -affinal 2.-male generation=1 geneal_distance=1 -affinal
3. syn	1.+male-1st-link generation=-1 geneal_distance=1 2.+male generation=-1 geneal_distance=1
4. corka	1.-male-1st-link generation=-1 geneal_distance=1 2.-male generation=-1 geneal_distance=1
5. siostra	1.-male-1st-link generation=0 geneal_distance=1 -affinal 2. -male generation=0 geneal_distance=1 -affinal 3. -male-1st-link -male generation=0 geneal_distance=1
6. brat	1. +male-1st-link generation=0 geneal_distance=1 -affinal 2. +male generation=0 geneal_distance=1 -affinal
7. siostra-stryjeczna	+male-1st-link +male-2nd-link -male generation=0
8. siostra-cioteczna	-male-2nd-link -male generation=0 geneal_distance=3
9. siostra-wujeczna	-male-1st-link +male-2nd-link -male generation=0 geneal_distance=3
10. brat-stryjeczna	+male-1st-link +male-2nd-link +male generation=0 geneal_distance=3
11. brat-cioteczna	-male-2nd-link +male generation=0 geneal_distance=3
12. brat-wujeczna	1. -male-1st-link +male-2nd-link +male generation=0 -affinal 2. -male-1st-link +male-2nd-link +male generation=0 geneal_distance=3
13. wuj	1. -male-1st-link +male generation=1 -affinal 2. -male-1st-link +male generation=1 geneal_distance=2 3. -male-1st-link +male-2nd-link generation=1 geneal_distance=2
14. stryj	1. +male-1st-link +male generation=1 geneal_distance=2 2. +male-1st-link +male-2nd-link generation=1 geneal_distance=2
15. ciotka	1. -male generation=1 geneal_distance=2 2. -male-2nd-link generation=1 geneal_distance=2
16. bratanek	1. +male-1st-link +male generation=-1 geneal_distance=2

ISBN: 0 08 043 438X

	2. +male-1st-link +male-2nd-link generation=-1 geneal_distance=2
17. siostrzeniec	-male-1st-link +male generation=-1
18. bratanica	+male-1st-link -male generation=-1
19. siostzenica	1. -male-1st-link -male generation=-1 geneal_distance=2
	2. -male-1st-link -male-2nd-link generation=-1 geneal_distance=2
20. dziad	1. +male generation=2
	2. +male-2nd-link generation=2
21. babka	1. -male generation=2
	2. -male-2nd-link generation=2
22. pradziad	+male generation=3
23. prababka	-male generation=3
24. wnuk	1. +male generation=-2
	2. +male-2nd-link generation=-2
25. wnuczka	1. -male generation=-2
	2. -male-2nd-link generation=-2
26. prawnuk	+male generation=-3
27. prawnuczka	-male generation=-3
28. maz	1. +male-1st-link geneal_distance=0
	2. +male geneal_distance=0
29. zona	1. -male-1st-link geneal_distance=0
	2.-male geneal_distance=0
30. œwiekr	1. +male-1st-link +male generation=1 +affinal
	2. +male-1st-link +male-2nd-link generation=1 +affinal
31. teœæ	1. -male-1st-link +male generation=1 +affinal
	2. -male-1st-link +male-2nd-link generation=1 +affinal
	3. -male-1st-link +male generation=1 geneal_distance=1
32. œwiekra	1. +male-1st-link -male generation=1 +affinal
	2. +male-1st-link -male-2nd-link generation=1 +affinal
	3. +male-1st-link -male generation=1 geneal_distance=1
33. teœciowa	1. -male-1st-link -male geneal_distance=1 +affinal
	2. -male-1st-link -male-2nd-link geneal_distance=1 +affinal
	3. -male-1st-link -male generation=1 +affinal
	4. -male-1st-link -male-2nd-link generation=1 +affinal
34. szwagier	1. +male generation=0 geneal_distance=1 +affinal
	2. +male-2nd-link generation=0 geneal_distance=1 +affinal
35. szwagrowa	1. +male-1st-link -male generation=0 +affinal
	2. -male generation=0 geneal_distance=1 +affinal
	3. -male-2nd-link generation=0 geneal_distance=1 +affinal
	4. +male-1st-link -male generation=0 geneal_distance=1

It may be noted that, apart from the more commonly used features in the European languages male, generation, and genealogical distance, Polish shares with Swedish the feature male-1st-link, absent, say, from languages like English and French. It does so in the contrasts between the pairs of kinterms:

wuj (*M*B) vs. *stryj* (*F*B),

wuj (*M*B) vs. *stryj* (*F*B),

bratanek (*B*s) vs. *siostrzeniec* (*S*s),

bratanica (*B*d) vs. *siostzenica* (*S*d).

This feature Polish shares also with some other Slavic languages (e.g. Bulgarian) which often (though less consistently than Swedish) distinguish between relatives on the male side and those on the female side (e.g. *M*B vs. *F*B in consanguineals or *W*B vs *H*B in affines). Unlike all other Slavic and European languages, Polish uses the feature male-2nd-link to be able to distinguish between the kinterm pairs:

siostra-stryjeczna (F*B*d) vs. *siostra-cioteczna* (F*S*d or M*S*d)

brat-stryjeczna (F*B*s) vs. *brat-cioteczna* (F*S*s or M*S*s)

Importantly, the analysis on Table 6 gives all alternative simplest kinterm definitions compliant with the six overall features that describe the whole corpus of kin data. As seen from the table, twelve of the kinterms have a unique meaning definition; two terms have four definitions, four kinterms three definitions, and the remaining seventeen two definitions.

We may wish to find the number of all "proper" componential models (i.e. where each kinterm has just one definition). If each kinterm definition can be used with anyone of the other kinterm definitions, the number of logically admissible componential models, Q, expressed in terms of the number of alternative kinterm definitions, N_I, can be found by the formula $Q = N_1 x N_2 x N_3 \ldots x N_m$; or, the number of all componential models equals the product of the number of definitions each kinterm has obtained. For Polish this means that we shall have $Q = 1^{12} x 2^{17} x 3^4 x 4^2 = 169\ 869\ 312$, that is, an almost astronomical number of kinship models! This multiplicity of solutions problem, looked from a theoretical perspective, has been noted in a famous article by Burling, 1964, but his stand has been disputed by other theorists as being just a theoretical speculation without important empirical implications. With KINSHIP we now have the possibility of finding all *empirically* valid models for a language (assuming a set of semantic features), and show that Burling was indeed correct in warning against the flood of alternative componential analyses. This problem however can be solved by adding further simplicity criteria, which are described in detail elsewhere (Pericliev and Valdes-Perez, submitted). We shall not enter into this problem here, but will only mention that these criteria would normally reduce an intolerably great number of alternatives to (nearly) unique analyses.

6. SOME FACILITIES OF *KINSHIP*

KINSHIP has a number of facilities that assist the user of the system in the study of kinship systems.

As regards the *input data*, KINSHIP maintains subroutines that check its "correctness". Thus, if some kintype is associated with more than one kinterm (which should not be the case with properly described kinship systems), then this cross-classification of kintypes is signalled by the system; also the system may check whether the kintypes listed are "syntactically correct" in the sense that they all comprise only sequences of the metalanguage (i.e. the primary

ISBN: 0 08 043 438X

genealogical relationships F, M, etc. and the auxiliary symbols used for denoting sex of ego, older/younger, etc.). Additionally, the user can query the system in a variety of ways as regards the input data, e.g. what kintypes are associated with a given kinterm; given a kintype, what is its corresponding kinterm, etc.

As regards semantic *features*, the user can obtain information as to what set of features are currently used by the system, and eventually select (for some reasons, e.g. psychological or social) a subset of these to accomplish the analysis. Also, one could retrieve what kintype/kinterm possesses what features.

The user can inquire about the *semantic contrasts* existing between a selected pair of kinterms or check what kinterms contrast with respect to a selected contrasting feature.

An important characteristic of the system is that it maintains *three styles of componential definitions* of kinterms. In the "fully redundant definition", each kinterm is defined with the values for all features necessary for the description of the whole dataset (if the componential model of a kinship system is stated in a tabular format, a fully redundant analysis will have *no* blank spaces, or cells, in the table). In a "non-redundand definition", each kinterm is defined with just those values which are both necessary and sufficient for discrimination of this kinterm from all the remaining kinterms; in all previous examples, we used this non-redundant style of componential definitions. The "partially redundant definition" stands in-between the other two styles in that it does not state in the meaning definition of a kinterm all values for all features, but may nevertheless contain values which are superfluous and hence of no import for the kinterms demarcation (for a detailed discussion of these definition styles, and examples, cf. Pericliev and Valdes-Perez (submitted)). We already mentioned the capability of the system to exhibit *all alternative componential analyses*, and the importance of this facility.

In the case of *failure to discriminate* between all kinterms, KINSHIP does not terminate its work, but rather proceeds smoothly, reporting the pairs of kinterms which it is unable to demarcate. This facility is quite useful in that it locates the problem and pinpoints where further human effort should be invested. However, alternatively, the system may try to automatically solve the problem by employing its mechanism for *inventing derived features*. To do so, KINSHIP combines its available features by the logical operations conjunction, disjunsction, implication and equivalence to form new, or derived, features.

Below we illustrate this mechanism on data from the Australian aboriginal language Gooniyandi (cf. McGregor, 1966). Running KINSHIP on this data, we found that, with the available features, a number of kinterm pairs cannot be distinguished. One such pair is:

jaja	MM MMS MMB MFSH WFM MBsW fds fdd
ngoomara	WMF

The system formed derived features, and found that the two kinterms are discriminable by the derived (binary) feature +/-[affinal-1st-link AND male-alter]:

jaja	-[affinal-1st-link AND male-alter]
ngoomara	+[affinal-1st-link AND male-alter]

ISBN: 0 08 043 438X

Indeed, the sole kintype of *ngoomara* has a first link which is +affinal (viz. W) *and* an alter which is +male (viz. F), while neither of the kintypes of *jaja* has *both* these features. The system also found that the same two kinterms are also distinguishable by the derived feature +/-[affinal-1st-link IMPLIES male-2nd-link]:

jaja	+[affinal-1st-link IMPLIES male-2nd-link]
ngoomara	-[affinal-1st-link IMPLIES male-2nd-link]

which can be easily checked to be true in all the kintypes of the corresponding kinterms.

Analogously, for another pair of terms:

ngawiji	FM FMS FMB fss fsd MMBW
jaminyi	MF MFB MFS WFF mds mdd

which were undistinguishable with the available primitive features are readily demarcated e.g. by the derived feature +/-[affinal-alter OR male-1st-link]:

ngawiji	+[affinal-alter OR male-1st-link]
jaminyi	-[affinal-alter OR male-1st-link]

since in all attendant kintypes of *ngawiji* the first link is either +male or the last link (=alter) is +affinal, and this is not the case in *either* of the kintypes of *jaminyi*.

7. CONCLUSION

KINSHIP has so far been applied to the analysis of more than 20 languages of different language families. The program's results on Seneca, as we saw, re-discover the meaning structure of terms as found by Lounsbury, which was a major advance in the understanding the Iroquois family relationships (for details, cf. Lounsbury 1964: 1079), showing at the same time the accuracy of his analysis. (It is worth noting that with a current redefinition of the feature «generation», to include counting «half generation», we managed to reduce the sufficient overall features for Seneca to 4.) The program was applied to unanalyzed kinship terminologies (e.g. Bulgarian), and has improved on previous componential analyses (e.g. English). Thus, for English KINSHIP has found more parsimonious analyses than those known in the literature, and the only ones that manage to give conjunctive (rather than occasionally, disjunctive) definitions of all kinterms, based on the detailed data of Goodenough, 1965 on Yankee kins. We have also tested componential analyses proposed by human kin analysts, revealing in some cases their logical inconsistency.

The need to formalize the task in a detail sufficient for computer implementation has led us to notice three different styles of writing componential definitions («fully redundant», «partially redundant», and «non-redundant»), depending on the presence/absence of redundant feature-values in a kinterm meaning description, which were not explicit in the previous literature. A further theoretical advance, made possible just because of the availability of our system, is related to the old problem of multiple solutions for kinship systems. Our system, with its capacity to produce alternatives, made conspicuous that not only theoretically, but practically, with the available features, the number of solutions can be immense (which we saw in the case of Polish) , which resulted in our proposing new simplicity criteria normally leading to unique, or just a few, componential analyses.

Also, we note that a part of the code of KINSHIP can directly be applied to other class discrimination tasks, whether in linguistics or outside of linguistics. We have so far applied it to the phonemic analysis of Russian and found a more parsimonious analysis than that proposed in the paper by Cherry, Halle and Jakobson, 1953.

We presume that all these facts taken together make quite a strong point in favour of machine discovery in linguistics. Further creative linguistic tasks should also be automated. However, even at this early stage, it seems clear that in tasks involving great bulks of data, such as componential analysis, a linguist teamed with a computer system has a much better chance of arriving at both logically consistent and (hopefully) significant results.

Acknowledgment. This work was supported by a grant #IRI-9421656 from the (USA) National Science Foundation, and by the NSF Disvision of Internaional Programs.

REFERENCES

Burling, R. (1964). Cognition and componential analysis: God's truth or hocus-pocus? *American Anthropologist* **66**, 20-28.

Cherry, C., M. Halle, and R. Jakobson (1953). Toward the logical description of languages in their phonemic aspect. *Language* **29**, 34-47

Goodenough, W. (1965). Yankee kinship terminology: a problem in componential analysis. In: *Formal Semantic Analysis*, (E. Hammel, (Ed.)). Special publication of *American Anthropologist* **67**, 259-287.

Kaluzin, L and E. Skorohodko (1963). Nekotorje zamecanija o leksiceskoj semantike (na materiale terminologii rodstva I svojstva). Issledovanija po strukturnoj tipologiii, (T. Molosnjaja (ed.)), 183-199. Izdatel'stvo Akademii Nauk SSSR: Moskva.

Kuznecov, A. (1974). On the typology of the semantic field of kinship terms. *Linguistics* **125**, 5-14.

Langley, P., H.Simon, G. Bradshaw, and J. Zytkow (1987). *Scientific Discovery: Computational Explorations of the Creative Processes.* MIT Press: Cambridge, Mass.

Leech, G. (1974). *Semantics.* Penguin: Harmondsworth.

Lounsbury, F. 1964. The structural analysis of kinship semantics. In: *Proc. of the 9th International Congress of Linguists*, (H. Lunt, (Ed.)), 1073-1090. Mouton and Co.: The Hague.

McGregor, W. (1966). Dyadic and polyadic kin terms in Gooniyandi. *Anthropological Linguistics* **38**, 216-247.

Pericliev, V. (1995). Empirical discovery in linguistics. In: *Working Notes of the Spring Symposium Series of AAAI: «Systematic Methods of Scientific Discovery»*, 68-73. AAAI Press, Palo Alto, California.

Pericliev, V. and R. E. Valdes-Perez (submitted). Automatic componential analysis of kinship semantics with a proposed solution to the problem of multiple models.

ISBN: 0 08 043 438X

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

Valdés-Pérez, R. E. (1996). Computer science research on scientific discovery. *Knowledge Engineering Review* **11**, 57-66.

ISBN: 0 08 043 438X