Historical-Comparative Reconstruction and Multilingual Lexica

James Kilbury[1], Katina Bontcheva[2]

[1] Computerlinguistik, Institut für Sprache und Information, Heinrich-Heine-Universität, Düsseldorf, Universitätsstr. 1, D-40225 Düsseldorf, Germany
kilbury@ling.uni-duesseldorf.de
http://web.phil-fak.uni-duesseldorf.de/~kilbury/
[2] Zieglerstr. 17, D-47058 Duisburg, Germany
katina.bontcheva@t-online.de

Abstract. This paper argues for the use of formal methods from historical-comparative reconstruction in the design of synchronic representations for multilingual lexica of genetically closely related languages. A model is discussed before an extended example with Slavic languages is given together with an implementation in DATR.

1 Introduction

A great deal of recent work in computational linguistics has been directed toward multilingual lexica (MLLs) for use in a wide variety of systems. Often it has been the case that particular MLLs have been developed for collections of languages dictated by external, often political, criteria. Thus, the project MULTEXT-East (cf. http://nl.ijs.si/ME/ or [9]) developed language resources for Bulgarian, Czech, Estonian, Hungarian, Romanian, and Slovene, as well as for English, the 'hub' language of the project.

Much less attention has been paid to the very special case of MLLs for genetically closely related languages. Even when genetically related languages have been dealt with, little effort has been made to capture common phonological and orthographic features of the languages.[1] For example, with regard to the West Germanic (WGmc) languages English, Dutch, and German, Cahill and Gazdar ([7]: 170, [8]: 11) point to such common features but do not show how to deal with them. In later work Tiberius and Cahill [18] introduce techniques for such correspondences but take no recourse to the comparative method we will employ in this paper. In quite a different vein Avgustinova and Uszkoreit [2] investigate shared morphosyntactic features of Slavic languages but do not address phonological questions.

[1] Cf. www.itri.brighton.ac.uk/projects/agile/ or [3] on AGILE.

2 Historical comparative reconstruction for multilingual lexica

It is, however, precisely the systematic sound correspondences which constitute the most salient feature of genetically closely related languages. While comparative linguistics of the 19th century began with the observation of shared *morphological* features, a solid basis for the methodology of *historical-comparative reconstruction* (HCR) was achieved only through the systematic investigation of sound correspondences, and in particular, through the methodological assumption of *regular sound change* by the Neogrammarians (cf. also [4]).

Once Saussure's distinction between *synchronic* and *diachronic* description had become established, investigators were wary lest they be accused of confusing the two. While caution was justified, the use of abstract morphophonemic representations *resembling* those postulated for earlier stages of a language frequently led to unfounded criticism that diachronic criteria had been employed in synchronic description. Aware of this danger, Bloomfield [5] warns that the representations in his treatment of Menomini morphophonemics *appear* to be those of a language stage arrived at through historical reconstruction, but in fact are justified in his synchronic description solely on the basis of morphophonemic alternations.

Indeed, the methods of internal reconstruction and morphophonemic representation are very similar but serve different goals. In an analogous way, a purely synchronic description of genetically closely related languages can take advantage of the historical method of HCR for the reconstruction of *proto forms* of an earlier, unattested source language. These methods are well known (cf. e.g. [12], [1]) and should be familiar to linguists.

3 A simple example from Germanic

In the case of the modern WGmc languages the phonological (as well as orthographic) similarities are immense. While the oldest attested stages of each language are employed in the historical reconstruction of the unattested source (Proto West Germanic), even the modern forms serve well for an exercise in the comparative method, as Anttila has shown in his textbook. Consider the following written forms:

English	**Dutch**	**German**
water	water	wasser
great	groot	gross
bite	bijt(en)	beissen/biss
hate	haat	hass
foot	voet	fuss
boat	boot	boot

Clearly, there is a systematic correspondence of English and Dutch /t/ in this postvocalic position to /s/ in (High) German. In the last set of cognates, the forms obviously are still related, but German has /t/ where we expect /s/. This turns out to

have arisen historically by the borrowing of *boot* into High German from Low German, which has the /t/ of the other languages. The exception, however, lessens in no way the importance of the *general and regular* correspondence seen in the other cognate sets. The methodology used here can be applied to other phonological and orthographic segments. In fact, we can derive both the phonological and orthographic forms of each cognate set from a single reconstructed representation, and this can be motivated without any reference to actual *historical* information.

4 Usefulness of the technique

Whether HCR can be useful for the synchronic description of a set of genetically related languages depends both on the transparency of the sound correspondences and on the existence of a large, culturally determined common lexical core. English, for example, has taken on a greater part of its vocabulary from Latin and French than have Dutch and German, so the applicability of the method may well be limited for these particular languages. We would hardly expect to profit from such techniques in a MLL including Spanish, Russian, and Gaelic, although all are Indo-European. In the case of the NGmc (Scandinavian) languages (especially Danish, Norwegian, and Swedish) the situation is quite different, as is reflected in the fact that speakers of one language normally do not attempt to learn actively another from the group. A similar situation can be seen in the Slavic, in particular, South Slavic, languages, which are characterized by extreme transparency of their interlinguistic phonological correspondences and by a vast core of common vocabulary arising from a common cultural heritage (in particular, the role of (Old) Church Slavonic for the development of the literary languages). Note that even here we find divergence in certain lexical layers (compare Russian *sobaka*, Bulgarian *kuče* and Serbo-Croatian *pas*, all 'dog'[2]) but much agreement in the learnèd vocabulary.

When such favorable conditions are present, the potential benefit of employing HCR in the synchronic design of a MLL is enormous and can be compared with the usefulness that morphophonemic techniques may have for the description of inflectional morphology (cf. [6]). In purely *theoretical* terms, the method allows linguistic generalizations about shared features to be stated. Probably of greater interest, however, are the *practical* advantages: the size of a MLL can be radically reduced, since a large number of forms can be derived from a single reconstructed representation rather than being stored separately. Moreover, perhaps the most important advantage is the promise of greater consistency and ease of updating and extending the MLL.

5 A formal model for HCR

The last-mentioned feature is typically cited as an argument in favor of inheritance hierarchies, and indeed, in our following example we use such hierarchies to model

[2] *Pes* exists also in Russian and Bulgarian but is rarely used.

the inheritance of features by Bulgarian and Macedonian from reconstructed South Slavic, and by the latter from Common Slavic. Like Tiberius and Cahill [18], we use the language DATR, which was designed specially for the representation of lexical information and has been employed in a wide variety of studies and applications (cf. [10]).

The prerequisites for our formal model of HCR are given in [13], in which Kay presents his analysis of Arabic morphology using *n*-tape *finite-state transducers* (FSTs; cf. [16] for a clear introduction). In accord with Kay's technique, interlingual sound correspondences simply constitute the alphabet of a FST, which has one tape for each daughter language and, if desired, may have an additional tape for each reconstructed language stage. The states of the FST can be used to model complementary distribution and thereby capture the positionally determined reflexes that arise through conditioned sound change.[3]

The model underlying our DATR implementation departs in some respects from the *n*-tape version but is equivalent to it. We have implemented conventional *2*-tape FSTs defining individual daughter languages as nodes in a hierarchy that inherit from a root representing their common, reconstructed features. DATR query paths correspond to individual reconstructed forms, while sequences returned as values by evaluations of node-path pairs represent the lexical forms of daughter languages.[4]

It is the *nonmonotonic* inheritance of DATR that allows us to capture the continuum of regularity, subregularity, and exceptions found in sound correspondences between languages. This provides a vehicle with which idiosyncratic forms like German *Boot* with /t/ or the various Slavic forms denoting 'dog' can be dealt with. Penn and Thomason [17] describe a closely related use of defaults in FSTs.

6 An extended example from Slavic

Our example assumes the following sound system for Slavic[5] (cf. [19] and [11]):

[3] In discussions with Kay at the 1987 EACL conference he agreed that the application of his technique to HCR would be obvious but had not considered the point himself. Kilbury [14] presented this formalization of HCR to an audience of historical linguists. In its functions our system resembles the Reconstruction Engine of Lowe and Mazaudon [15], but the latter has no theoretical basis involving FSTs and draws no connection to MLLs.

[4] At least two alternative designs are conceivable: (1) one could model reconstructed forms as DATR nodes and paths as languages, so that returned values represent individual reflexes, e.g. `RYBA:<bul> = r i b a`, or (2) one could instead model lexemes as nodes and paths as languages, again, with values as reflexes, e.g. `PISCIS:<bul> = r i b a`. In each of the three designs, we clearly see DATR as a *functional* formalism.

[5] Notes on transliteration: cf. Table 3. Transliteration of the Cyrillic characters.

vowels		**consonants and some clusters**
i y u		p b t d (tj dj) k g
ï ü		s z sh zh x
e o		c ch dzh
ë ö		m n (nj)
ä a		l (lj)
		r (rj)
		w j

Certain vowel correspondences which are the basis of our example are given in the following table for GEN(eral)SL(avic)[6] as well as RUS(sian), BUL(garian), and MAC(edonian):

Table 1. Vowel correspondences for GENSL, RUS, BUL, MAC ('-' denotes zero)

GENSL	*i	*ï	*e	*ë	*ä	*a	*y	*ö	*o	*ü	*u
RUS	i	e/ï	e	q	e	a	y	u	o	o/-	u
BUL	i	e/-	e	e	e/q	a	i	ü	o	ü/-	u
MAC	i	e/-	e	e	e	a	i	a	o	o/-	u

Note that for all reflexes in the daughter languages, we have here employed a Roman transliteration of the Cyrillic orthographies. These phonological correspondences are illustrated in the following cognate sets, which give the reflexes as well as a reconstructed form and a gloss:

Table 2. Cognate sets for our example

GENSL	**RUS**	**BUL**	**MAC**	**gloss**
ögül	ugol	ügül	agol	angle
zvün	zvon	zvün	zvon	ringing/clang
dostöp	dostup	dostüp	dostap	access
ryba	ryba	riba	riba	fish
vremë	vremq	vreme	vreme	time/weather
pëtï	pqtï	pet	pet	five
snä´g	sneg	snqg	sneg	snow
xlä´b	xleb	xlqb	leb	bread
dïnï	denï	den	den	day
bedstvie	bedstvie	bedstvie	betstvie	disaster
sbor	sbor	sbor	zbor	sum
teatr	teatr	teatür	teatar	theater
dualism	dualizm	dualizüm	dualizam	dualism

[6] We have invented this term to emphasize that our reconstructed forms differ from those of Old Church Slavonic or Common Slavic and that they make no historical claims.

gõrd	gorod	grad	grad	town
gõls	golos	glas	glas	voice
bãr´g	bereg	brqg	breg	strand/bank
zhãlza	zheleza	zhleza	zhleza	gland

7 Implementation in DATR

The DATR encoding of our description requires a set of *variable declarations* (cf. [10]: 187), which capture natural sets of segments:

```
# vars $segm:    p b t d k g x f v s z c m n r l w j h
                 i e a o u ï ë ä ö y ã õ @ q.
# vars $cons:    p b t d k g x f v s z c m n r l w j.
# vars $mute:    p t k x f s c .
# vars $voiced:  b d g z .
# vars $vow:     i e q a ü o u @.
# vars $sonor:   m n r l .
# vars $liquida: r l .
```

The rest of the DATR *theory* (i.e., program) consists of node definitions for the central hierarchy of FSTs (cf. [10]: 191-193 for the encoding of FSTs). In addition to nodes for the three daughter languages, we have one node each for the reconstructed stages GENSL and S(outh)SL(avic). All the nodes inherit from a default identity transducer ELSE, which simply maps a path into the identical sequence.

```
Else:
  <> ==
  <$X> == $X "<>" .

GENSL:
  <> == Else
  <ä> == "<e>"                        % ь >e
  <ï $cons> == "<e $cons>"            % ь >e
  <$segm ´> == "<$segm>"              % stress sign removal
  <* $segm> == "<$segm>"              % word boundaries removal
  <$segm *> == "<$segm>" .
```

```
RUS:
  <> == GENSL
  <ë> == "<q>"                          % ѧ > я
  <ö> == "<u>"                          % ѫ > y
  <ü> == "<o>"                          % ъ > o
  <õ $liquida> == "<o $liquida o>"      % *tORt/tOLt
  <ã $liquida> == "<e $liquida e>" .    % *tERt/tELt

SSL:
  <> == GENSL
  <ë> == "<e>"                          % ѧ > e
  <y> == "<i>"                          % ы > и
  <$cons ï *> == "<$cons>" .  % loss of word final palatalness

BUL:
  <> == SSL
  <ä ´> == "<q>"                        % stressed ь > я
  <ö> == "<ü>"                          % ѫ > ъ
  <$cons $sonor *> == "<$cons ü $sonor>" % ъ epenthesis
  <õ $liquida> == "<$liquida a>"        % *tORt/tOLt
  <ã $liquida> == "<$liquida ä>"  .     % *tERt/tELt

MAC:
  <> == SSL
  <ö> == "<a>"                          % ѫ > a
  <ü> == "<o>"                          % ъ > o
  <d $mute> == "<t $mute>"              % regressive assimilation[7]
  <s $voiced> == "<z $voiced>"          %
  <* x $segm> == "<$segm>"              % elision of word initial x
  <$cons $sonor *> == "<$cons a $sonor>" % a epenthesis
  <õ $liquida> == "<$liquida a>"        % *tORt/tOLt
  <ã $liquida> == "<$liquida e>"  .     % *tERt/tELt
```

By exploiting the notational possibilities of DATR we have encoded the transducer for each language in a single node. Alternatively, each transducer could be represented as a subnetwork of nodes, where additional nodes are introduced to implement states of a FST that capture the environments preceding and following a positionally conditioned correspondence. Suitable *hide* and *show* declarations allow us to generate a dump with exactly the set of lexical correspondences given in the table above.

[7] The orthography of Macedonian reflects regressive assimilation of mute and voiced consonants but never the devoicing in word final position. Here we have two instances of regressive assimilation.

8 Conclusions

Our approach relies entirely on well-established methods of historical linguistics coupled with those of computational linguistics. We note that all the orthographic representations for the forms in the individual daughter languages can be derived from a single reconstructed representation. In a simple extension, the phonological representations, which can in turn be used for speech synthesis, may be derived from the same reconstructions, so that a high degree of economy is achieved. While maintaining our approach, we can replace our linear segmental transcriptions with hierarchical phonological representations like those of Cahill and Gazdar ([8]: 20). We have established a systematic basis that permits the development of a much larger lexicon, and moreover, the group of included languages can easily be extended to other SSlav languages (Serbo-Croatian and Slovene) and ESlav languages (Byelorussian and Ukrainian).

Most of all, we want to emphasize that our proposals are complementary to and compatible with the work that has been done by others for the syntax and morphology of genetically closely related languages as well as with other studies using DATR. It therefore utilizes and extends well-developed resources that are already available.

References

1. Anttila, R.: Historical and Comparative Linguistics. 2nd revised edition. John Benjamins Publishing Company, Amsterdam & Philadelphia (1989)
2. Avgustinova, T., Uszkoreit, H.: An ontology of systematic relations for a shared grammar of Slavic. Proceedings of COLING 2000 28-34
3. Bateman, J., Teich, E., Kruijff-Korbayová, I., Kruijff, G.-J., Sharoff, S., Skoumalová, H.: Resources for multilingual text generation in three Slavic languages. In: Gavrilidou, M., Carayannis, G., Markantonatou, S., Piperidis, S., Stainhaouer, G. (eds.): Proceedings of Second International Conference on Language Resources and Evaluation ELRA (2000) 1763-1768
4. Bloomfield, L.: A note on sound change. Language 4 (1928) 99-100
5. Bloomfield, L.: Menomini morphophonemics. Travaux du Cercle Linguistique de Prague 8 (1939) 105-115
6. Bontcheva, K., Kilbury, J.: An inheritance-based description of Bulgarian noun inflection. Proceedings of the Workshop on Balkan Language Resources and Tools. Thessaloniki (2003)
7. Cahill, L. J., Gazdar, G.: Multilingual lexicons for related languages. Proceedings of the 2nd DTI Language Engineering Conference (1995) 169-176
8. Cahill, L. J., Gazdar, G.: The POLYLEX architecture: multilingual lexicons for related languages. Traitement Automatique des Langues 40 (1999) 7-25
9. Erjavec, T., Ide, N., Petkevic, V., Véronis, J.: Multext-East: multilingual text tools and corpora for central and eastern European languages. Proceedings of the First European TELRI Seminar: Language Resources for Language Technology (1996) 87-98

10.Evans, R., Gazdar, G.: DATR: a language for lexical knowledge representation. Computational Linguistics 22 (1996) 167-216
11.Friedman, V.: Macedonian. In: Comrie, B., Corbett, G. G. (eds.): The Slavonic Languages. Routledge, London (1993) 249-305
12.Hoenigswald, H. M.: Language Change and Linguistic Reconstruction. University of Chicago Press, Chicago London (1960)
13.Kay, M.: Nonconcatenative finite-state morphology. Proceedings of the 3rd Conference of the European Chapter of the Association for Computational Linguistics (1987) 2-10
14.Kilbury, J.: Automaton theory and the formalization of historical-comparative reconstruction. Unpublished paper presented at the 13th International Conference on Historical Linguistics, 10-17 August 1997 in Düsseldorf
15.Lowe, J. B., Mazaudon, M.: The Reconstruction Engine: a computer implementation of the comparative method. Computational Linguistics 20 (1994) 381-417
16.Sproat, R.: Morphology and Computation. MIT Press, Cambridge, Mass. (1992)
17.Penn, G., Thomason, R.: Default finite state machines and finite state phonology. Proceedings of the 1st SIGPHON Workshop (1994) 33-42
18.Tiberius, C., Cahill, L.: Incorporating metaphonemes in a multilingual lexicon. Proceedings of COLING 2000 1126-1130
19.Townsend, C. E., Janda, L. A.: Common and Comparative Slavic: Phonology and Inflection. Slavica Publishers, Columbus, Ohio (1996)

Appendix

Table 3. Transliteration of the Cyrillic characters

Cyrillic	Latin		Cyrillic	Latin		Cyrillic	Latin
а	a		й	j		у	u
б	b		к	k		ф	f
в	v		л	l		х	x
г	g		м	m		ц	c
д	d		н	n		ч	ch
е	e		о	o		ш	sh
ѧ	ë		ѫ	ö		щ	shh
ь	ä		о (in tort/tolt)	õ		ъ	ü
е (in tert/telt)	ã		п	p		ы	y
ж	zh		р	r		ь	ï
з	z		с	s		ю	@
и	i		т	t		я	q

This abstract Latin transliteration for Slavic was created by Katina Bontcheva in order to capture adequately both orthography and pronunciation (stress has not yet been fully dealt with). It differs from the ISO standard in several ways:

- *jat* and *nasal o, e* are represented with ä, ö, ë
- o, e in tORt/tOLt and tERt/tELt groups are represented with õ, ã
- *ž, č, š* are represented with the digraphs *zh, ch, sh* because of restrictions on the diacritics imposed by the DATR
- for the same reason ъ and ь are transcribed with *ü* and *ï* (instead of " and ')
- digraphs were avoided for the transcription of *ю* and *я* and @ and *q* used instead in order to keep as close as possible to Cyrillic orthography
- щ is shh .