Tools and Resources for Bibliographic Data in Cross-Linguistic Research

Harald Hammarström
harald2@chalmers.se

June 15, 2009

Bibliographic Data in Typology

Today's Talk:

- Desiderata: A Practical Approach
- Resources: Existing bibliographical databases
- Tools: A simple method for auto-annotation of bibliographical references

Bibliographical Desiderata for Typologists

- A bibliography of all relevant research articles?

  Too large to be feasible!

- A bibliography of *descriptive* materials of the languages of the world?
  - Language documentation and description is, and has been, an extremely decentralised activity
  - There are over 500 bibliographies of descriptive materials in printed form, e.g.

    P. Newman 1996 *Hausa and the Chadic Language Family: A Bibliography* Köln: Köppe [African Linguistic Bibliographies 6].
  - Now, in the digital age, it may be possible to combine all previous listings and make a near-complete continually updated database!

Goal

A bibliography website

- Browsing
- Searching
- Downloading
- Updating
  - Wiki?
  - Benevolent dictatorship?
- New items subscription
- Etc.

Publications with Descriptive Data

BDP = A bibliographic reference to a publication with descriptive/documentational data on a lesser-known language

- I have a database of about 14 000 BDP:s
- Empirically, BDP:s are of *two prototypical kinds*:

  Individual Descriptions: Dammann, Ernst 1957 *Studien zum Kwangali: Grammatik, Texte, Glossar*, Hamburg: Cram, de Gruyter & Co. [Abhandlungen aus dem Gebiet der Auslandskunde / Reihe B, Völkerkunde, Kulturgeschichte und Sprachen 35]

  Group Descriptions: Donald C. Laycock 1968 *Languages of the Lumi Subdistrict*, Oceanic Linguistics VII(1):36-66

- In my database, ca 28% is of the group kind overall (though number varies a lot across areas)

Practical Annotation Proposal: Focus

- **Language-id** for individual BDP:s

  ISO 639-3 code (from which location, speaker number etc. is derivable separately)

- **Group-id** for group BDP:s

  Group-id:s could be any name with geographical, genealogical or other inspiration which is *equated with a set of language-id:s* separately from the annotation of the language entry.

(This is a specialised form of the doculect-langoid scheme of Good & Hendryx 2006, Good & Cysouw 2007 where special prominence is given to the kind of langoids that BDP:s typically instantiate.)

Practical Annotation Proposal: Type

Type: According to the following relatively uncontroversial hieracharchy:

- (full-length) descriptive grammar
- grammar sketch
- description of some element of grammar (i.e. noun class system, verb morphology etc)
- phonological description
- dictionary
- text (collection)
- wordlist
- document with meta-information about the language (i.e., where spoken, non-intelligibility to other languages etc.)
- note on unpublished manuscripts or people engaged in studying the language

Summing Up: Practical Desiderata

Defining the goal:

- References to descriptive language data is a delineable class
- Annotation of focus and type meets basic search needs

Getting to the goal:

- Collecting all refs is feasible?
- Doing the annotation is feasible?

Important Resources I know of

	# Refs	Contents	Area	Cov.	Annot.
EBALL	49 442	Everything	Africa	Full	L & T
Fabre	ca 45 600	Everything	S America	Full	L
Hammarström	14 075	Descriptive data	World	85%?	T
EVA	ca 11 700	Everything	World	?	L & T
SIL	14 826	Everything	World	95%?	L & T
SILPNG	ca 13 110	Everything	Papua	Full	L & T

Availability?

- Availability?
  - SIL and EVAMPG are queryable on the net
  - SILPNG and Fabre are on the Web in the raw text form
  - Hammarström will be available for download
  - EBALL is queryable on the net

Some more figures from Maho (EBALL)

- Figures

  31 549 Annotated for language

  11 335 Grammar, morphology, syntax
  6 279 Phonetics/phonology (incl. tonology)
  3 777 Dictionaries, longer wordlists
  3 026 Grammar introductions, overviews

- Details on annotation + samples: `http://goto.glocalnet.net/maho/eball.html`
  online querying: `http://sumale.vjf.cnrs.fr/Biblio/index.html`

Summing Up: Existing Resources

- The bulk of collecting references into electronic form has largely been done already
- Likely these collections can be used for benevolent purposes
- A lot of annotation work is remaining nevertheless

Automatic Annotation of Focus

Given: A database of the world's languages

Input: A bibliographical reference to a work with descriptive language data (= a BDP) of (at least one of) the language in the database

Desired output: The identification of which language(s) is described in the bibliographical reference

Unfortunately, the problem is not simply a clean database lookup!

Example

Dammann, Ernst 1957 *Studien zum Kwangali: Grammatik, Texte, Glossar*, Hamburg: Cram, de Gruyter & Co. [Abhandlungen aus dem Gebiet der Auslandskunde / Reihe B, Völkerkunde, Kulturgeschichte und Sprachen 35]

- This reference happens to be written in German. In general, the metalanguage could be any language.
- This reference happens to describe a Namibian-Angolan language called Kwangali, ISO 639-3 *kwn*
- The task is to automatically infer this
  - using a database of the world's languages and/or databases of other annotated bibliographical entries
  - but without humanly tuned thresholds

Motivation

- There are about 7 000 languages in the world
- Language description, i.e., producing a phonological description, grammatical description, wordlist, dictionary, text collection or the like, of these 7 000 languages has been on-going on a larger scale since about 200 years.
- This process is fully de-centralized, and at present there is no database over which languages of the world have been described, which have not, and which have partial descriptions already produced
- We are conducting a large-scale project of listing all published descriptive work on the languages of the world, especially lesser-known languages.

Similar Work?

- Annotation of bibliographical entries with language appears to be a previously untargeted problem
- However, it is a special case of a more general Information Extraction (IE) problem:
  - There is a set of natural language objects O
  - There is a fixed set of categories C
  - Each object in O belong to zero or more categories

  The special case we are considering here is such that:
  - An object contains only a small amount of text $\sim$ 100 words
  - The language of objects in O varies
  - $|C|$ is large, i.e., there are many classes $\sim$ 7 000
  - $|C(o)|$ is small for most objects $o \in O$, i.e., most objects belong to very few categories, typically ~ 1
  - Most objects $o \in O$ contain a few tokens that near-uniquely identifies $C(o)$

Specifics: World Language Database

- The Ethnologue `http://www.ethnologue.com` is a database that aims to catalogue all the known living languages of the world.
- Each language is given a unique three-letter identifier, a canonical name and a set of variant and/or dialect names.
- Example:

  Canonical name: Kwangali
  ISO 639-3: kwn
  Alternative names: {Kwangali, Shisambyu, Cuangar, Sambio, Kwangari, Kwangare, Sambyu, Sikwangali, Sambiu, Kwangali, Rukwangali}.

Specifics: Languages and Language Names

- 7 299 languages
- 42 768 language name tokens
- 39 419 unique name strings
- It is not yet well-understood how "complete" this language name database is. However:
  - 100 randomly chosen bibliographical entries contained 104 language names in the title.
  - 43 of these names (41.3%) existed in the database as written.
  - 66 (63.5%) existed in the database allowing for variation in spelling

Free Annotated Databases

- Training of a classifier ('language annotator') in a supervised framework, requires a set of annotated entries with a distribution similar to the set of entries to be annotated.
- Two such databases which can be freely accessed

  WALS: The bibliography for the *World Atlas of Language Structures* `http://www.wals.info/`: 5633 entries annotated to 2053 languages.

  MPI/EVA: The library catalogue of the Max Planck Institute for Evolution Anthropology `http://biblio.eva.mpg.de/` (May 2006): 7266 entries annotated to 2246 languages.
- For training and development, we used both databases put together.
- The two together, duplicates removed: 8584 entries annotated to 2799 languages.

Data to be annotated

- Currently 7804 entries need to be annotated (no overlap with the joint WALS-MPI/EVA database)
- The (meta-)languages of the entries are English, German, French, Spanish, Portuguese, Russian, Dutch, Italian, Chinese, Indonesian, Thai, Turkish, Persian, Arabic, Urdu, Nepali, Hindi, Georgian, Japanese, Swedish, Norwegian, Danish, Finnish and Bulgarian
- From the 7 804 entries, 100 were randomly selected and humanly annotated to form a test set.
- This test set was not used in the development at all, and was kept totally fresh for the final tests.

Experiments

Naive Union Lookup: Each word in the title is looked up as a possible language name in the world language database and the output is the union of all answers to the look-ups.

Term Weight Lookup: Each word is given a weight according to the number of unique-id:s it is associated with in the training data. Based on these weights, the words of the title are split into two groups; informative and non-informative words. The output is the union of the look-up:s of the informative words in the world language database.

Term Weight Lookup with Group Disambiguation: As above, except that names of genealogical (sub-)groups and country names that occur in the title are used for narrowing down the result.

Further Notes

Spelling: Enrich database of language names which machine generated but realistic language name spelling variation

Accuracy: Measure two kinds of accuracy:

- **Perfect Accuracy:** The gold standard set of languages and algorithm output have to match exactly
- **Sum Accuracy:** $\frac{|\{X(e) \cap e_c\}|}{|e_c \cup X(e)|}$: The overlap between the gold standard and algorithm output (match with score between 0 and 1)

Spelling Normalization

#	Substition Reg. Exp.	Replacement	Comment
1.	\'\‘\^\~\"	''	diacritics truncated
2.	[qk](?=[ei])	qu	k-sound before soft vowel
3.	k(?=[aou]|$)|q(?=[ao])	c	k-sound before hard vowe
4.	oo|ou|oe	u	oo, ou, oe to u
5.	[hgo]?u(?=[aouei]|$)	w	hu-sound before hard vow
6.	((?:[^aouei]*[aouei] [^aouei]*)+?)		
	(?:an$\|ana$|ano$\|o$)	\1a	an? to a
7.	eca$	ec	eca to ec
8.	tsch|tx|tj	ch	tsch, tx to ch
9.	dsch|dj	j	dsch, dj to j
10.	x(?=i)	sh	x before i to sh
11.	i(?=[aouei])	y	i before a vowel to y
12.	ern$\|i?sche?$	''	final sche, ern removed
13.	([a-z])\1	\1	remove doublets
14.	[bdgv]	b/p,d/t,g/k,v/f	devoice b, d, g, v
15.	[oe]	o/u,e/i	lower vowels

Naive Union Lookup

$$NUL(e) = \cup_{w \in Words(e_t)} LN(w)$$

Anne Gwenaïélle Fabre 2002 *Étude du Samba Leko, parler d'Allani (Cameroun du Nord, Famille Adamawa)*, PhD Thesis, Université de Paris III – Sorbonne Nouvelle

$Words(e_t)$	$LN(Words(e_t))$	$Words(e_t)$	$LN(Words(e_t))$
etude	$\{\}$	cameroun	$\{\}$
du	$\{dux\}$	du	$\{dux\}$
samba	$\{ndi, ccg, smx\}$	nord	$\{\}$
leko	$\{ndi, lse, lec\}$	famille	$\{\}$
parler	$\{\}$	adamawa	$\{\}$
d'allani	$\{\}$		

- $NUL(e) = \{ndi, lse, smx, dux, lec, ccg\}$, but the correct classification is $e_c = \{ndi\}$.
- Accuracy on test set $PA_{NUL}(A) \approx 0.15$ and $SA_{NUL}(A) \approx 0.21$.

Naive Lookup is Too Naive

- Clearly, we cannot guess blindly which word(s) in the title indicate the target language!
- BUT we can exploit some domain specific properties:
  - A title of a publication in language description typically contains
    1. One or few words with very precise information on the target language(s), namely the name of the language(s)
    2. A number of words which recur throughout many titles, such as ’a’, ’grammar’, etc.
  - Most of the languages of the world are poorly described, there are only a few, if any, publications with original descriptive data.

Term Weight Lookup

- Inspired by *t f-id f*
- Measure the informativeness of a word *w*: $WC(w)$ = the number of distinct codes associated with *w* in the training data or Ethnologue database
- At which point (above which value?) of informativeness do we get a near-unique language name rather than a relatively ubiquitous non-informative word?
- Luckily, we are assuming that there are only those two kinds of words, and that at least one near-unique language will appear.
- Thus, if we cluster the values into two clusters, the two categories are likely to emerge nicely.
- The simplest kind of clustering of scalar values into two clusters is to sort the values and put the border where the relative increase is the highest.

Example

W. M. Rule 1977 *A Comparative Study of the Foe, Huli and Pole Languages of Papua New Guinea*, University of Sydney, Australia [Oceania Linguistic Monographs 20]

	foe	pole	huli	papua	guinea	comparative	new	study	languages	and	a	the	of
$WC(w)$	1	2	3	57	106	110	145	176	418	1001	1101	1169	1482
Rel.Inc.	1.0	2.0	1.5	19.0	1.86	1.04	1.32	1.21	2.38	2.39	1.10	1.06	1.27

- The highest relative increase is 19.0 between Huli and Papua
- Thus, Foe, Pole and Huli are deemed near-unique and the rest non-informative.
- In this example, the three near-unique identifiers are correctly singled out

Term Weight Lookup

- Denote $SIG_{WC}(e_t)$ the group of most informative words in a title e_t. We can restrict lookup only to them:

$$TWL(e) = \cup_{w \in SIG_{WC}(e_t)} LN(w)$$

- In the example above, $TWL(e_t)$ is $\{fli, kjy, foi, hui\}$ which is almost correct, containing only a spurious [fli] because Huli is also an alternative name for Fali in Cameroon, nowhere near Papua New Guinea.
- The resulting accuracies jump up to $PA_{TWL}(A) \approx 0.57$ and $SA_{TWL}(A) \approx 0.73$.

Adding Group Disambiguation

- We know that a large number of entries contain a “group name”, i.e., the name of a country, region of genealogical (sub-)group in addition to a near-unique language name.
- Since group names will naturally tend to be associated with many codes, they will sorted into the non-informative camp with the *TWL*-method, and thus ignored.
- This is unfortunate, because such group names can serve to disambiguate inherent small ambivalences among near-unique language names, as in the case of Huli above.
- Group names are not like language names. They are much fewer, they are typically longer (often multi-word), and they exhibit less spelling variation.
- A database of group names is easily generated from the Ethnologue: 3 202 groups

TWL with Group Disambiguation

- The non-significant words of a title is searched for matching group names. The set of languages denoted by a group name is denoted $L(g)$ with $L(g) = C$ if g is not a group name found in the database.

$$TWG(e) = (\cup_{w \in SIG_{WC}(e_t)} LN(w)) \cap_{g \in (Words(e_t) \setminus SIG_{WC}(e_t))} L(g)$$

- We get slight improvements in accuracy $PA_{TWG}(A) \approx 0.59$ and $SA_{TWG}(A) \approx 0.74$. The corresponding accuracies with spelling variation enabled are $PA_{TWG}(A) \approx 0.64$ and $SA_{TWG}(A) \approx 0.77$.

Discussion

	PA	SA
NUL	0.15	0.21
TWL	0.57	0.73
TWL_S	0.61	0.74
TWI	0.55	0.70
TWI_S	0.59	0.71
TWG	0.59	0.74
TWG_S	0.64	0.77

- All scores conform to expected intuitions and motivations.
- The key step beyond naive lookup is the usage of term weighting (and the fact the we were able to do this without a threshold or the like).
- In the future, it appears fruitful to look more closely at automatic extraction of groups from annotated data.

Summing Up: Automatic Annotation

- There is an algorithm for language-id annotation of BDP:s with
  - useful accuracy
  - little or no human work
- Can automatic annotation of type also be done?

Summing Up: Overall

A website with a comprehensive and annotated database of bibliographical references to lesser known languages (BDPs) is not too far away

- I am committed to working on it
- Help always needed
- What are your ideas for update management?
- How does one create a *functioning* collaborative resource?

The End

Thank You for Listening!