

Quality control for digitized dictionaries

Paul Rodrigues, David Zajic, Tim
Buckwalter, Michael Maxwell,
C. Anton Rytting

prodrigues@casl.umd.edu

Who are we?

- University of Maryland Center for Advanced Study of Language (CASL)
 - We write language grammars and clean up dictionaries for the US Government.
 - The languages have limited *computational* resources.
 - They may be under-described in the *linguistic* literature.
 - Dialectal variation, based on poorly collected data, theoretical disagreements between linguists, etc.
 - Data and XML are errorful.

What languages?

Language	Grammar	Dictionary
Bengali	✓	
Urdu	✓	✓
Pashto	✓	✓
Iraqi Arabic		✓
Yemeni Arabic		✓
Modern Standard Arabic		✓
Dhivehi (Maldivian)	✓	?

After this?

LANGUAGE RESEARCH IN SERVICE TO THE NATION

Quality control for digitized dictionaries

Sections

1	Errors in digitized dictionaries and how they are introduced
2	Automatic Anomaly Detection
(2.5)	(Grammar/Parser Testing)
3	Error Fixing with the Dictionary Manipulation Language
4	Future Work (BRIDGE, Crowdsourcing)

Structure Problems

- Dictionaries use fonts and spacing for structure.
 - Improper indentation, **bolding**, *italicizing*, or underlining changes the structure of the electronic representation.

Structure Problems - Example

ارقام *irqām'* N.M. (ped.) writing putting
, down in black and white ارقام کرنا *irqam*
kar'nā v.tr. pen ; write ارقام ہونا *irqām' ho'na* v.i.
be penned ; be written *arqām* N.M. PL.
figures ; numbers items [A ~ SING. رقم]

Qureshi, B.A., Abdul Haq. 2003. Standard 21st Century Dictionary.
Educational Publishing House, Delhi, India


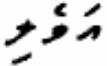

Structure Problems - Example

ارقام *irqām'* N.M. (ped.) writing putting
, down in black and white ارقام کرنا *irqam*
kar'nā v.tr. pen ; write ارقام ہونا *irqām' ho'na* v.l.
be penned ; be written ارقام N.M. PL.
figures ; numbers items [A ~ SING. رقم]

Missing Calligraphy – About 200 occurrences in dictionary


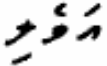

Qureshi, B.A., Abdul Haq. 2003. Standard 21st Century Dictionary.
Educational Publishing House, Delhi, India

Structure Problems - Example 2

- 205 AVIHI (avissek)  n. 1. kind of dragonfly
2. a former intercalary asterism [Skt. Abhijit]
- 206 AVELI  See avali
- 207 AS (ahek)  n. 1. horse; knight (at chess) [S. as, Skt. aśva]
2. row of knots: a.jahanii, ties knots; as bai, double knot: faḷi as, slipknot
3. See as-duuni

Reynolds, C. 2003. A Maldivian Dictionary.
RoutledgeCurzon. New York.

Structure Problems - Example 2

- 205 AVIHI (avissek)  n. 1. kind of dragonfly
2. a former intercalary asterism [Skt. Abhijit]
- 206 AVELI  See avalī
- 207 AS (ahék)  n. 1. horse; knight (at chess) [S. as, Skt. aśva]
2. row of knots: a.jahanii, ties knots; as bai, double knot: faḷi as, slipknot
3. See as-duuni

No typographical difference or separation between Dhivehi pronunciations and English words.

Reynolds, C. 2003. A Maldivian Dictionary.
RoutledgeCurzon. New York.

Content Problems

- Spelling, Typographical errors anywhere in entry
 - in headword or examples (L1!)
 - in definition (L2!)
- Pronunciation errors.
- Grammar errors
 - On examples (L1!)
 - On definitions (L2!)
- Lexicographer ‘issues’
 - Semantic issues with sense analysis.
 - Inconsistencies between authors.

Content Problems - Example

مضيف *muḍiif* f. -a p. -iin host, air host.
مضيفة *muḍiifah* air hostess, stewardess.

دكتور *dxtwr*

دكتور *daxtoor* (common var. *taxtoor*) p. *daxaatir*, *taxaatir*
doctor (medical doctor and Ph.D.) f.
-ah p. -aat. كشف عليّ التختور وقال لازم

Qafisheh, H.A. 2000. NTC's Yemeni Arabic-English Dictionary. NTC Publishing Group. Chicago.

LANGUAGE RESEARCH IN SERVICE TO THE NATION

Content Problems - Example

مضيف *muḍiif* f. -a p. -iin host, air host.
مضيفة *muḍiifah* air hostess, stewardess.

دخاتور *dxtwr*

دكتور *daxtoor* (common var. *taxtoor*) p. *daxaatir*, *taxaatir*
doctor (medical doctor and Ph.D.) f.
-ah p. -aat. كشف عليّ التختور وقال لازم

Inconsistent phonetic representation of Feminine Suffix

Qafisheh, H.A. 2000. NTC's Yemeni Arabic-English Dictionary. NTC Publishing Group. Chicago.

LANGUAGE RESEARCH IN SERVICE TO THE NATION

What about errors that can *get introduced* by Digitalization?

Electronic-origin

- General Spelling Errors
- General Typographical Errors

Paper Origin w/Manual Electronic Transfer

*All Electronic-origin errors,
plus....*

- Bad Handwriting and Ambiguity introduced by the Calligrapher
- Typesetting to format
 - Whitespace misalignments
 - Fonts
- OCR Errors

Lexicographers or data entry clerks may not be native speakers of (either) language.

What about OCR?

- No “mixed” language model. (Half this line will be English, and half will be O’zbek, but “That’s ok!”)
 - In fact, often no language model at all for one of the languages.
- No dictionary specific language model.
 - Abbrevs. s.a. Adv., N.
 - T y p o g r a p h y fonts, *italics*, underline, **bold**)
- Handwriting slips of the pen, crammed text.

OCR Problems - Illustration

ارقام *irqām'* N.M. (ped.) writing putting
, down in black and white ارقام كونا *irqam*
kar'nā v.tr. pen ; write ارقام هونا *irqam' ho'na* v.l.
be penned ; be written ارقام N.M. PL.
figures ; numbers items [A ~ SING. رقم]

Just the handwriting...

Qureshi, B.A., Abdul Haq. 2003. Standard 21st Century Dictionary.
Educational Publishing House, Delhi, India

Finding Structural Errors

- We tried language modeling.
 - Create a language model off flattened structure, test flattened structure against that model,
 - SRI Language Modeling Toolkit trained, tested on flattened dictionary structure.
 - Sorted by perplexity.
 - Odd combinations of strings appear at the top.
- Unfortunately we lose structure.

Finding Content Errors (1/4)

- Anomalous spelling and pronunciation detection
 - Dictionaries provide parallel corpus of native orthographic representation and transliterations.
 - We used alignment and the confidence of statistical transliteration algorithms
 - Align at the character level via Giza++
 - Sort by likelihood, and examine low probability alignments.

Finding Content Errors (2/4)

- Disambiguating code points that were conflated.
- Low probability of dominant alignment may indicate ambiguous code point
 - Hex E7: 81% h-slash, 14% h
- Use alignment character to disambiguate
 - h-slash: dochashmi he
 - h: choti he (Urdu)

Finding Content Errors (3/4)

- QC via Finite State Morphological Parsers
 - Several of the languages have ongoing morphological documentation projects.
 - The morphological grammars are written and researched collaboratively between linguists and computational linguists.
 - Descriptive Grammar / Formal Grammar
 - Integrating the parser with a lexicon creates a testable book.
 - Errors can be found in the grammar, in the lexicon, or from the original printed dictionary. (this *has* happened)

Finding Content Errors (4/4)

- What about grammar? (e.g. Example sentences)
 - Not as easy.
 - We're not writing syntactic grammars.
 - n-grams are not enough.
 - Maybe corpus search with word replaceability.

We found the errors... now fix them!

- Nothing beats a language expert.
- Our algorithms highlight possible errors, but we require people to *make* the change.
 - We do this for accuracy. We don't want to *introduce* errors.
 - (We don't trust our algorithms yet.)

Language Experts

- Look for major error patterns that can be modified en masse by a programmer.
- After the major modifications, spend hours each day comparing digital lexicon to original dictionary.
- Additionally filter through the output created by unsupervised algorithms.
- Make lists of errors.
 - Dictionaries have lots of errors. In a team of computer scientists and linguists, the computer scientists become the bottleneck!

Keeping a record

- We're usually not the first team to touch the dictionary, and we may not be the last.
- Sometimes mistakes were made that we'd like to undo. Sometimes WE make mistakes we'd like to undo.
- Making changes directly to the XML or to a relational database that outputs XML would limit future workers to a chronological undo.

Dictionary Manipulation Language

- Dictionary Manipulation Language (DML) was created to give editorial control to the language expert.
 - Simple programming language designed for XML dictionary manipulation.
- All updates are stored as DML commands whether generated manually or by automatically
- Selective undo, with only local effect

Dictionary Manipulation Language

- Commands create nodes, move nodes, set attributes and modify underlying text

```
create internal FORM after 19406 F
move 19413 under F
set attribute F TYPE variant
sub text 89730 "understnad" "understand"
```

Language Expert Workflow

- Language expert works independently of computer scientist:
 - discovers problem in output XML file
 - writes new DML commands to fix problem
 - runs all DML commands against source XML file
 - checks if problem is fixed in output XML file
- In work on Yemeni dictionary:
 - 250 problems solved by computer scientist
 - 3600 problems solved by language experts

Summary

- Errors can be introduced at any stage, including attempts to fix other problems.
- Errors can be found using anomaly detection.
- We still have to fix errors by using smart language experts.
- An XML edit history/changelog allows us to easily integrate the language experts into the process, and allows a non-chronological undo.

Problems/Future Work

- How can we reduce the errors?
 - For the content errors, is there a way to tell if these are errors or *variation*? e.g. regional pronunciation differences.
- How can we mark output of statistical algorithms as definite errors?
- Can we group the errors and fix them automatically?
- How can we expand the coverage of a dictionary, update dictionaries to keep them current? (Fast, Cheap, and Automatic?)

Back to Paper-origin → Electronic Transfer

- Structured Text OCR
 - BRIDGE enables rapid structured importation of paper dictionaries using trained tagging that relies on consistency of lexical entry layout. Ma et. al, (2003)
 - Up to 93% accuracy on tagging Karagol-Ayan et. al (2006)
 - could influence the structure of lexical items in a hierarchical lexicon.

Crowd Sourcing

- What about Mechanical Turk?
 - Copyright Issues with displaying dictionary pages?
 - Ensuring a contributor's fluency? (Snow et al., 2008)
 - Variation?
 - Would take a LONG TIME for Less-Resourced Languages Irvine and Klementiev (2010), Maxwell and Hughes (2006)
 - Still relatively unresearched.

References

Karagol-Ayan, B., Doermann, D., and Weinberg, A. Adaptive Transformation-based Learning for Improving Dictionary Tagging. Proceedings of the 11th Conference on European Chapter of the Association for Computational Linguistics, pages 257-264, April 2006.

Ma, H., Karagol-Ayan, B., Doermann, D., Oard, D., and Wang, J. Tagging and Parsing of Bilingual Dictionary. Technical Report: LAMP-TR-106/CFAR-TR-991/CS-TR-4529/UMIACS-TR-2003-97, University of Maryland, College Park, September 2003.

Irvine, A., Klementiev, A. 2010. Using Mechanical Turk to Annotate Lexicons for Less Commonly Used Languages. Proceedings of NAACL 2010 Workshop on creating Speech and Language Data with Amazon's Mechanical Turk. Los Angeles.

Maxwell, M. and Hughes, B. 2006. Frontiers in Linguistic Annotation for Lower-Density Languages. Proceedings of the ACL Workshop on Frontiers in Linguistically Annotated Corpora. p29-37. Sydney.

Snow, R., O'Connor, B., Jurafsky, D., Ng, A. 2008. Cheap and Fast—But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. Proceedings of EMNLP 2008.