“Did You Mean...?” and Dictionary Repair: from Science to Engineering

Michael Maxwell, presented by Petra Bradley
(mmaxwell@casl.umd.edu, pbradley@casl.umd.edu)
University of Maryland, College Park MD 20742 USA

AMTA 2016

29 October 2016

Two problems with electronic (XML) dictionaries

- How can you look up a word? Especially when...
  - You can't spell gud
  - There's noise in the environment
  - You can't hear well
  - It's in a foreign language
- What happens when the dictionary has an error?
  - Missing information, information in the wrong field, wrongly structured fields
  - Errors range from typos to missing pages' worth of data
  - CASL has worked with dozens of XML dictionaries, mostly from government sources; no dictionary has been without errors
  - Errors are pervasive: ∼10,000 errors in one Urdu dictionary (with ∼50,000 entries)

CASL's solutions

- "Did You Mean...?" (DYM)
  Smart fuzzy lookup. Three versions:
  - Basic DYM: Spell correction only
  - Morphologically Aware DYM (MADYM): Spell correction + morphological parsing
  - Cross-language DYMs: Basic or MADYM + lookup in multiple dialect/ language/ script dictionaries
    - Waziri Pashto: lookup in 1902 Waziri dictionary, plus backoff lookup in modern Pashto dictionary via sound changes
    - French-in-Arabic: lookup of Arabic script queries in Moroccan Arabic dictionary *and* in French dictionary (with Arabic morphological parsing)
- Dictionary repair tools
  - ADALT discovers anomalies: rare structures, data that doesn't "fit the mold"
  - VELMA is an editor specialized for dictionaries and similar documents

Tools

ADALT + VELMA and DYM are *tools*; neither replaces human labor

- J.C.R. Licklider, 1960 *Man-Computer Symbiosis*: “Computing machines can do readily, well, and rapidly many things that are difficult or impossible for man, and men can do readily and well, though not rapidly, many things that are difficult or impossible for computers. That suggests that a symbiotic cooperation, if successful in integrating the positive characteristics of men and computers, would be of great value.”
- ADALT and VELMA complement human abilities in error-finding
- DYM enables humans to better understand critical foreign language texts by doing smart fuzzy lookup

Two problems with electronic (XML) dictionaries

- How can you look up a word? Especially when...
  - You can't spell gud
  - There's noise in the environment
  - You can't hear well
  - It's in a foreign language
  - *"Did You Mean...?"*
- What happens when the dictionary has an error?
  - Missing information, information in the wrong field, wrongly structured fields
  - Errors range from typos to missing pages' worth of data
  - CASL has worked with dozens of XML dictionaries, mostly from government sources; no dictionary has been without errors
  - Errors are pervasive: ∼10,000 errors in one Urdu dictionary (with ∼50,000 entries)

How we got there: DYMs

- 2007: Dr. Anton Rytting (CASL) suggests need for spell correction for dictionary lookup by Arabic language analysts
- 2008: First “Did You Mean…?” for Georgetown Iraqi Arabic dictionary.
- 2009: Second DYM: Urdu, overwhelmingly positive response.
- Later: Additional Arabic dictionaries, gazetteers, 10,001 Arabic names; Pashto, Russian, Ukrainian, Somali, Persian (Farsi), Swahili…
- 2012-2013: Dr. Corey Miller adds Persian morphological parser to Persian DYM to produce first Morphologically Aware DYM (MADYM)
- Later: MADYMs for Arabic, Somali, Swahili, Korean; DYMs for Chinese (pinyin), Dhivehi, Punjabi, Portuguese (Open Street Maps for Rio de Janeiro)
- 2016:
  - DYM Toolkit for building (part basic of) DYMs for new languages
  - DYM Platforms for deploying basic DYMs and MADYMs

Electronic Dictionary Lookup

Ordinary dictionary lookup: type in a word, get back definition(s) *if* you spelled it correctly, and you chose the dictionary citation form

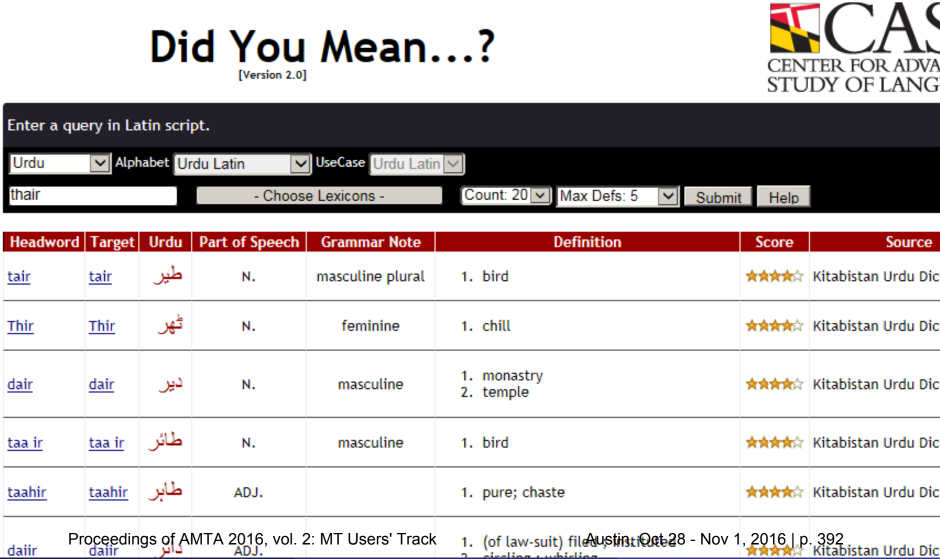
Wildcard lookup: For each letter you're uncertain about, type in a '*'
If you hear German [rat], type *ra**.
But this will give you unwanted words: *Rah*, *Rap*, *Rank*, *rar*, *rau*, *Rat*, and *Rad*–whereas only the last two are likely to be what you want.

Regular expression lookup: If you think it might end in a 'd' or a 't', type *ra[d|t]*. Works if you understand regular expressions, and if you know where the mistakes are likely to be, and if you remember what the possibilities are.

DYM: Builds in the knowledge of regular expressions, likely mistakes, possible substitutions (and deletions and insertions), and *how likely* a particular mistake is; and automagically applies this knowledge everywhere in the word where a mistake is possible. Type in *rat*, get back *Rat* and *Rad* (but not the others).

How does a DYM work? outside the box

Urdu Romanization. th = aspirated dental, t = unaspirated dental, Th = aspirated retroflex.



Did You Mean...?

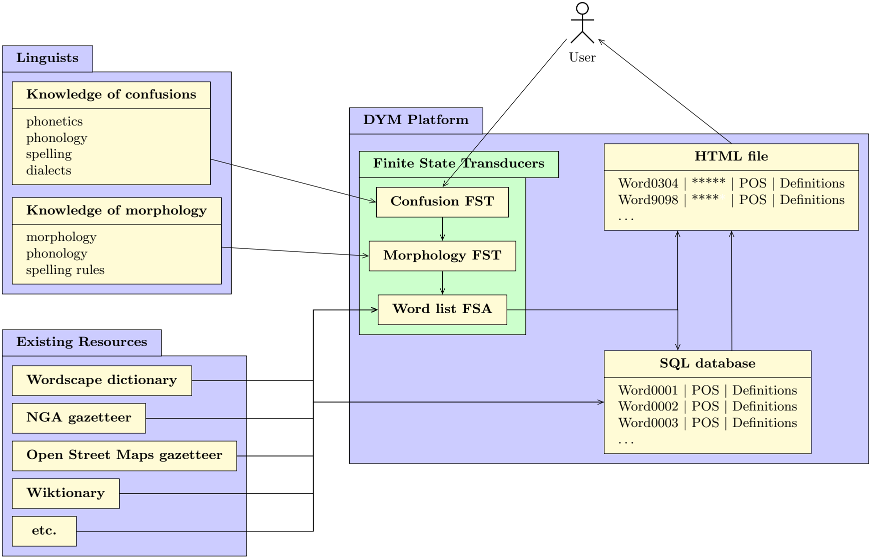
[Version 2.0]

CAS
CENTER FOR ADVA
STUDY OF LANG

Enter a query in Latin script.

Headword	Target	Urdu	Part of Speech	Grammar Note	Definition	Score	Source
tair	tair	طير	N.	masculine plural	1. bird	★★★★☆	Kitabistan Urdu Dic
Thir	Thir	ٹھر	N.	feminine	1. chill	★★★★☆	Kitabistan Urdu Dic
dair	dair	دير	N.	masculine	1. monastry 2. temple	★★★★☆	Kitabistan Urdu Dic
taa ir	taa ir	طائر	N.	masculine	1. bird	★★★★☆	Kitabistan Urdu Dic
taahir	taahir	طاہر	ADJ.		1. pure; chaste	★★★★☆	Kitabistan Urdu Dic
daiir	daiir	دائر	ADJ.		1. (of law-suit) filed ; instituted 2. circling ; whirling	★★★★☆	Kitabistan Urdu Dic

How does a DYM work? inside the box



User

Linguists

Knowledge of confusions

phonetics
phonology
spelling
dialects

Knowledge of morphology

morphology
phonology
spelling rules

Existing Resources

Wordscape dictionary

NGA gazetteer

Open Street Maps gazetteer

Wiktionary

etc.

DYM Platform

Finite State Transducers

Confusion FST

Morphology FST

Word list FSA

HTML file

Word0304 | ***** | POS | Definitions
Word9098 | **** | POS | Definitions
. . .

SQL database

Word0001 | POS | Definitions
Word0002 | POS | Definitions
Word0003 | POS | Definitions
. . .

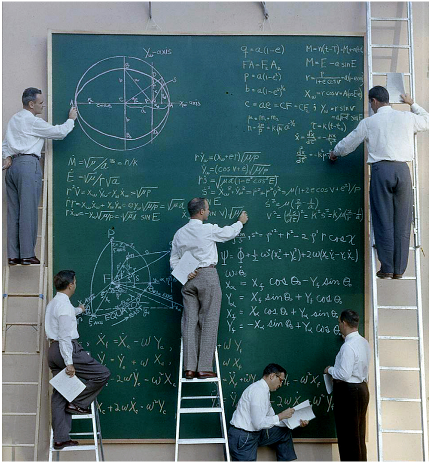
But that’s so complicated!

CASL tools to build DYMs:

- DYM Toolkit to incorporate the linguists’ knowledge of confusions
  - DYM Toolkit enables end users (e.g. language analysts) to build confusion matrices for their confusions.
  - The toolkit can also be used to build confusion matrices for dialects, non-standard spelling systems, native speaker errors.
    - In conjunction with a transliterator, this can be used for Romanized scripts (Arabeze, Pinglish/ Farslish,...)
- Dictionaries, gazetteers often exist (but stay tuned for part 2 of this talk)
- DYM Platform takes care of the software infrastructure

But that's so complicated!

Morphological parsing...



...is still rocket science (and optional).

DYM software

- Platforms 1 and 2, with DYMs, submitted as deliverable (includes documentation and about ten DYMs); USG has free use.
  - Also available with consulting/ training.
  - DYM Toolkit included.
- What do I need to build a new DYM?
  - Dictionary or other lexical resource (e.g. gazetteer)
    If it needs cleanup…stay tuned!
  - Confusion matrix
    - Available for non-native listener confusions for ten languages
    - …or language analysts can build their own with the provided Toolkit
  - …or talk to us about building a new one.

DYM software (cont'd)

Planned work

- Cross-language DYMs (like Waziri Pashto, French-in-Arabic)
  - CASL has built such DYMs, but current Platforms do not support them

    …future work.
- Stand-alone DYMs (without network connection/ server)
  - This could also be used to rapidly field stand-alone dictionaries, even without a 'real' confusion matrix.
- Coming soon to your cellphone?

Two problems with electronic (XML) dictionaries

- How can you look up a word? Especially when…
  - You can't spell gud
  - There's noise in the environment
  - You can't hear well
  - It's in a foreign language
- What happens when the dictionary has an error?
  - Missing information, information in the wrong field, wrongly structured fields
  - Errors range from typos to missing pages' worth of data
  - CASL has worked with dozens of XML dictionaries, mostly from government sources; no dictionary has been without errors
  - Errors are pervasive: ∼10,000 errors in one Urdu dictionary (with ∼50,000 entries)
  - *ADALT and VELMA*

VELMA: dictionary editor

- VELMA = Visual Environment for Lexicography and MAchine learning

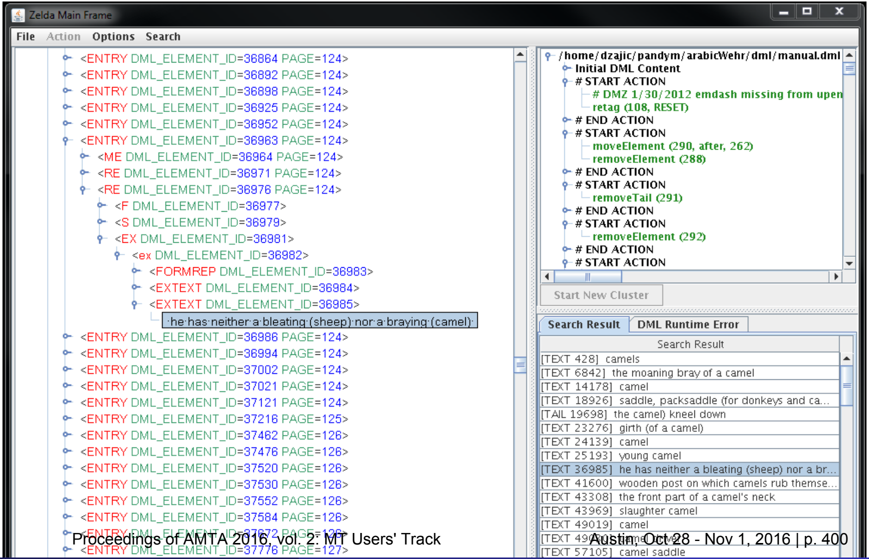
New version = Zelda

- VELMA is designed for editing of existing dictionaries (not building new ones)
  - Allows easy manipulation of existing XML nodes and text...
  - ...meaning Ordinary Working Lexicographers (not computer scientists)* can modify dictionaries.
  - Integrated with ADALT.
  - Uses ADALT output to suggest a workflow for linguists/ lexicographers.

    *Ok, you need to learn about XML...

VELMA in pictures

Dictionary file on left, actions performed upper right pane; search results (or ADALT output) lower right pane (here: search for ‘camel’).



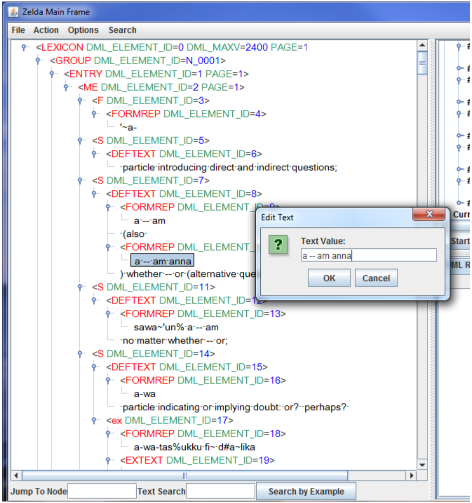
```
<ENTRY DML_ELEMENT_ID=36864 PAGE=124>
<ENTRY DML_ELEMENT_ID=36892 PAGE=124>
<ENTRY DML_ELEMENT_ID=36898 PAGE=124>
<ENTRY DML_ELEMENT_ID=36925 PAGE=124>
<ENTRY DML_ELEMENT_ID=36952 PAGE=124>
<ENTRY DML_ELEMENT_ID=36963 PAGE=124>
  <ME DML_ELEMENT_ID=36964 PAGE=124>
  <RE DML_ELEMENT_ID=36971 PAGE=124>
  <RE DML_ELEMENT_ID=36976 PAGE=124>
    <F DML_ELEMENT_ID=36977>
    <S DML_ELEMENT_ID=36979>
    <EX DML_ELEMENT_ID=36981>
      <ex DML_ELEMENT_ID=36982>
        <FORMREP DML_ELEMENT_ID=36983>
        <EXTEXT DML_ELEMENT_ID=36984>
        <EXTEXT DML_ELEMENT_ID=36985>
          he has neither a bleating (sheep) nor a braying (camel)
<ENTRY DML_ELEMENT_ID=36986 PAGE=124>
<ENTRY DML_ELEMENT_ID=36994 PAGE=124>
<ENTRY DML_ELEMENT_ID=37002 PAGE=124>
<ENTRY DML_ELEMENT_ID=37021 PAGE=124>
<ENTRY DML_ELEMENT_ID=37121 PAGE=124>
<ENTRY DML_ELEMENT_ID=37216 PAGE=125>
<ENTRY DML_ELEMENT_ID=37462 PAGE=126>
<ENTRY DML_ELEMENT_ID=37476 PAGE=126>
<ENTRY DML_ELEMENT_ID=37520 PAGE=126>
<ENTRY DML_ELEMENT_ID=37530 PAGE=126>
<ENTRY DML_ELEMENT_ID=37552 PAGE=126>
<ENTRY DML_ELEMENT_ID=37584 PAGE=126>
<ENTRY DML_ELEMENT_ID=37672 PAGE=126>
<ENTRY DML_ELEMENT_ID=37776 PAGE=127>
```

```
/home/dzajic/pandym/arabicWehr/dml/manual.dml
  Initial DML Content
  # START ACTION
    # DMZ 1/30/2012 emdash missing from upen
    retag (108, RESET)
  # END ACTION
  # START ACTION
    moveElement (290, after, 262)
    removeElement (288)
  # END ACTION
  # START ACTION
    removeTail (291)
  # END ACTION
  # START ACTION
    removeElement (292)
  # END ACTION
  # START ACTION
```

```
Search Result
[TEXT 428]  camels
[TEXT 6842]  the moaning bray of a camel
[TEXT 14178]  camel
[TEXT 18926]  saddle, packsaddle (for donkeys and ca...
[TAIL 19698]  the camel) kneel down
[TEXT 23276]  girth (of a camel)
[TEXT 24139]  camel
[TEXT 25193]  young camel
[TEXT 36985]  he has neither a bleating (sheep) nor a br...
[TEXT 41600]  wooden post on which camels rub themse...
[TEXT 43308]  the front part of a camel's neck
[TEXT 43969]  slaughter camel
[TEXT 49019]  camel
[TEXT 490[illegible]]  camel driver
[TEXT 57105]  camel saddle
```

VELMA in pictures

Supports typical editor operations: search and replace, drag-and-drop, change text (or XML tags), undo (of *any* action, not just most recent)



```
<LEXICON DML_ELEMENT_ID=0 DML_MAXV=2400 PAGE=1
  <GROUP DML_ELEMENT_ID=N_0001>
    <ENTRY DML_ELEMENT_ID=1 PAGE=1>
      <ME DML_ELEMENT_ID=2 PAGE=1>
        <F DML_ELEMENT_ID=3>
          <FORMREP DML_ELEMENT_ID=4>
            '~a-
        <S DML_ELEMENT_ID=5>
          <DEFTEXT DML_ELEMENT_ID=6>
            ·particle·introducing·direct·and·indirect·questions;
        <S DML_ELEMENT_ID=7>
          <DEFTEXT DML_ELEMENT_ID=8>
            <FORMREP DML_ELEMENT_ID=9>
              a·--·am
            ·(also·
            <FORMREP DML_ELEMENT_
              a·--·am·anna
            )·whether·--·or·(alternative·que
        <S DML_ELEMENT_ID=11>
          <DEFTEXT DML_ELEMENT_ID=12>
            <FORMREP DML_ELEMENT_ID=13>
              sawa~'un%·a·--·am
            ·no·matter·whether·--·or;
        <S DML_ELEMENT_ID=14>
          <DEFTEXT DML_ELEMENT_ID=15>
            <FORMREP DML_ELEMENT_ID=16>
              a-wa
            ·particle·indicating·or·implying·doubt:·or?··perhaps?·
          <ex DML_ELEMENT_ID=17>
            <FORMREP DML_ELEMENT_ID=18>
              a-wa-tas%ukku·fi~·d#a~lika
            <EXTEXT DML_ELEMENT_ID=19>
```

Edit Text

Text Value: a -- am anna

So what is this ADALT stuff?

How do you find errors in the electronic equivalent of hundreds of pages of a dictionary?

- Errors come in two types:
  - Frequent: Some common structure is mis-represented throughout
    - Easy for humans to notice at least some of them.
    - ...but finding all instances of an error type can be hard!
  - Rare: One-offs, typos
    - Needle-in-a-haystack
    - Hard for humans to find these...
    - ...especially when you don't know what kinds of needles there are!

Finding those needles in a haystack

You need...

CASL’s metal detector

ADALT (Automatic Detection of Anomalies in Lexicographic Text)

- ADALT uses machine learning (AI) to find rare structures
- These are anomalies; some anomalies are errors, some are simply rare structures.
- Examples:
  - (Spanish, but based on actual anomaly in an Urdu dictionary):

    raro *adj.* (Usage: rare)

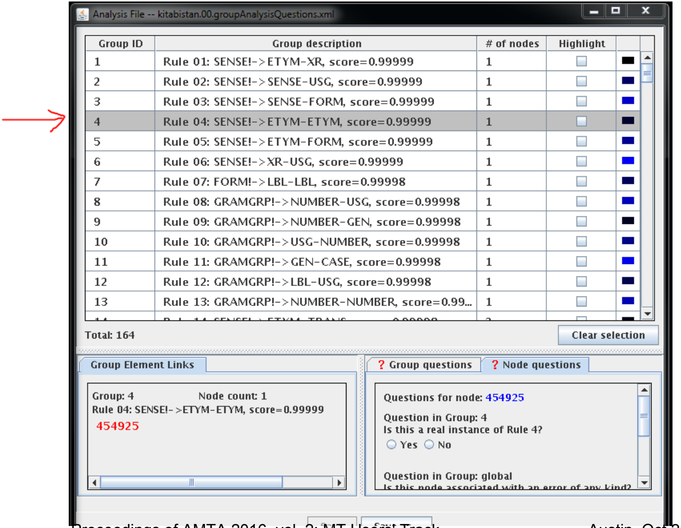
    Missing definition!
  - (English, but similar anomalies found in Urdu dictionary):

    colo(u)r ……

    Prevents lookup: user will never search for this spelling!
  - (Illustrative; actual errors would be in XML)
- ADALT builds models of structures found in a particular dictionary, and uses heuristics to find structures which are rare within those models.
- It therefore doesn’t know (and doesn’t need to be told) what is ‘correct’.

Example of ADALT usage: the metal detector

The metal detector has found anomalies; user chooses one to work on (a single lexeme that has two etymologies):



Group ID	Group description	# of nodes	Highlight	
1	Rule 01: SENSE!->ETYM-XR, score=0.99999	1	☐	■
2	Rule 02: SENSE!->SENSE-USG, score=0.99999	1	☐	■
3	Rule 03: SENSE!->SENSE-FORM, score=0.99999	1	☐	■
4	Rule 04: SENSE!->ETYM-ETYM, score=0.99999	1	☐	■
5	Rule 05: SENSE!->ETYM-FORM, score=0.99999	1	☐	■
6	Rule 06: SENSE!->XR-USG, score=0.99999	1	☐	■
7	Rule 07: FORM!->LBL-LBL, score=0.99998	1	☐	■
8	Rule 08: GRAMGRP!->NUMBER-USG, score=0.99998	1	☐	■
9	Rule 09: GRAMGRP!->NUMBER-GEN, score=0.99998	1	☐	■
10	Rule 10: GRAMGRP!->USG-NUMBER, score=0.99998	1	☐	■
11	Rule 11: GRAMGRP!->GEN-CASE, score=0.99998	1	☐	■
12	Rule 12: GRAMGRP!->LBL-USG, score=0.99998	1	☐	■
13	Rule 13: GRAMGRP!->NUMBER-NUMBER, score=0.99...	1	☐	■
14	Rule 14: SENSE!->ETYM-TRANS, score=0.99998	2	☐	■

Total: 164

Group: 4 Node count: 1
Rule 04: SENSE!->ETYM-ETYM, score=0.99999
454925

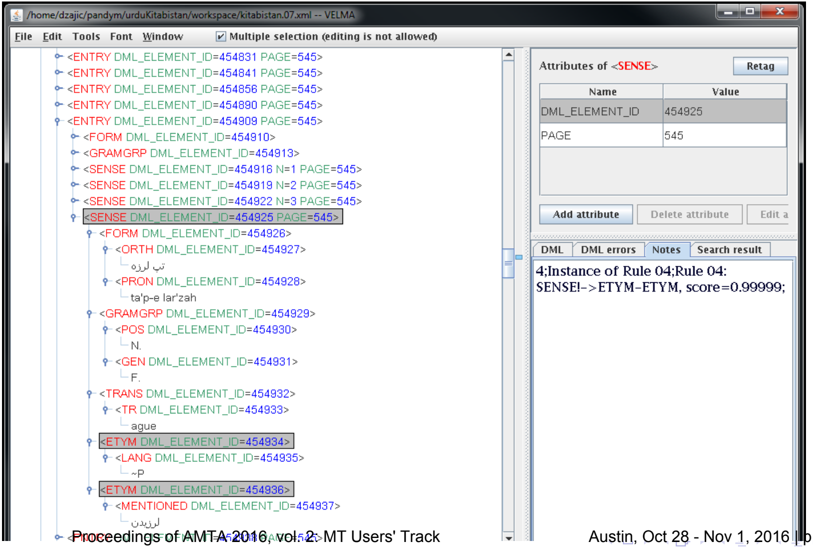
Questions for node: 454925

Question in Group: 4
Is this a real instance of Rule 4?
○ Yes ○ No

Question in Group: global
Is this node associated with an error of any kind?

Example of ADALT usage: The needle is found!

VELMA jumps to that anomaly; the first ‘etymology’ contains just a language tag (Persian), the second lacks a language tag.


/home/dzajic/pandym/urduKitabistan/workspace/kitabistan.07.xml -- VELMA
File Edit Tools Font Window
☑ Multiple selection (editing is not allowed)
Retag
Add attribute
Delete attribute
Edit a
DML
DML errors
Notes
Search result

```
<ENTRY DML_ELEMENT_ID=454831 PAGE=545>
<ENTRY DML_ELEMENT_ID=454841 PAGE=545>
<ENTRY DML_ELEMENT_ID=454856 PAGE=545>
<ENTRY DML_ELEMENT_ID=454890 PAGE=545>
<ENTRY DML_ELEMENT_ID=454909 PAGE=545>
  <FORM DML_ELEMENT_ID=454910>
  <GRAMGRP DML_ELEMENT_ID=454913>
  <SENSE DML_ELEMENT_ID=454916 N=1 PAGE=545>
  <SENSE DML_ELEMENT_ID=454919 N=2 PAGE=545>
  <SENSE DML_ELEMENT_ID=454922 N=3 PAGE=545>
  <SENSE DML_ELEMENT_ID=454925 PAGE=545>
    <FORM DML_ELEMENT_ID=454926>
      <ORTH DML_ELEMENT_ID=454927>
        تپ لرزہ
      <PRON DML_ELEMENT_ID=454928>
        ta'p-e lar'zah
    <GRAMGRP DML_ELEMENT_ID=454929>
      <POS DML_ELEMENT_ID=454930>
        N.
      <GEN DML_ELEMENT_ID=454931>
        F.
    <TRANS DML_ELEMENT_ID=454932>
      <TR DML_ELEMENT_ID=454933>
        ague
    <ETYM DML_ELEMENT_ID=454934>
      <LANG DML_ELEMENT_ID=454935>
        ~P
    <ETYM DML_ELEMENT_ID=454936>
      <MENTIONED DML_ELEMENT_ID=454937>
        لرزیدن
```

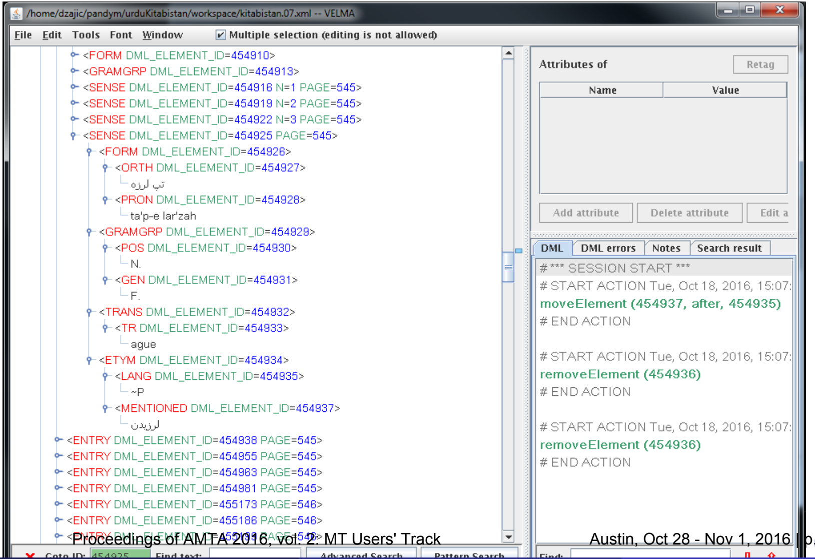
Attributes of <SENSE>

Name	Value
DML_ELEMENT_ID	454925
PAGE	545

4;Instance of Rule 04;Rule 04: SENSE!->ETYM-ETYM, score=0.99999;

Example of ADALT usage: grab the needle

The user uses VELMA to remove the needle from the haystack (merge the two etymologies)



```
    <FORM DML_ELEMENT_ID=454910>
    <GRAMGRP DML_ELEMENT_ID=454913>
    <SENSE DML_ELEMENT_ID=454916 N=1 PAGE=545>
    <SENSE DML_ELEMENT_ID=454919 N=2 PAGE=545>
    <SENSE DML_ELEMENT_ID=454922 N=3 PAGE=545>
    <SENSE DML_ELEMENT_ID=454925 PAGE=545>
      <FORM DML_ELEMENT_ID=454926>
        <ORTH DML_ELEMENT_ID=454927>
          تپ لرزہ
        <PRON DML_ELEMENT_ID=454928>
          ta'p-e lar'zah
      <GRAMGRP DML_ELEMENT_ID=454929>
        <POS DML_ELEMENT_ID=454930>
          N.
        <GEN DML_ELEMENT_ID=454931>
          F.
      <TRANS DML_ELEMENT_ID=454932>
        <TR DML_ELEMENT_ID=454933>
          ague
      <ETYM DML_ELEMENT_ID=454934>
        <LANG DML_ELEMENT_ID=454935>
          ~P
        <MENTIONED DML_ELEMENT_ID=454937>
          لرزیدن
  <ENTRY DML_ELEMENT_ID=454938 PAGE=545>
  <ENTRY DML_ELEMENT_ID=454955 PAGE=545>
  <ENTRY DML_ELEMENT_ID=454963 PAGE=545>
  <ENTRY DML_ELEMENT_ID=454981 PAGE=545>
  <ENTRY DML_ELEMENT_ID=455173 PAGE=546>
  <ENTRY DML_ELEMENT_ID=455186 PAGE=546>
  <ENTRY DML_ELEMENT_ID=[illegible]>
```

Name	Value

Goto ID: 454925

```
# *** SESSION START ***
# START ACTION Tue, Oct 18, 2016, 15:07:
moveElement (454937, after, 454935)
# END ACTION

# START ACTION Tue, Oct 18, 2016, 15:07:
removeElement (454936)
# END ACTION

# START ACTION Tue, Oct 18, 2016, 15:07:
removeElement (454936)
# END ACTION
```

…about those frequent errors

Latest version of VELMA (Zelda!) provides “query-by-example” (QBE) and “edit-by-example” (EBE)

- If the user notices multiple of instances of some structural error…
  - Dozens might be evident on inspection
  - …and hundreds more might be lurking.
- …then use QBE to find all instances of that structure (or similar structures).
- Once the user has verified that QBE finds the desired erroneous structures, correct one, then use EBE to correct others in the same way.

ADALT and VELMA software

- Mature: CASL has used this (or previous versions) on about a dozen dictionaries.
- VELMA is preferable to text editors/ Perl/ version control systems because…
  - Users are less likely to make mistakes
  - …and if you do make a mistake, it can be undone six months from now without messing up everything done between now and then.
  - Provides audit trail of changes.
  - Users report working faster.
- Submitted as deliverable (includes documentation); USG has free use.
  - Also available with consulting/ training.
- What can I do with this?
  - Import/ clean up dictionaries
  - Import/ clean up gazetteer data
  - Potentially useful for other kinds of semi-structured text data

Summary

- DYMs and dictionary repair started as CASL research projects less than ten years ago.
- Many useful tools and resources (cleaned up dictionaries, DYMs) came out of the work along the way.
- Dictionary repair and building DYMs are now (mostly) engineering.
  - Fielding a new DYM can be done in a few days.
  - Repairing a dictionary can take a few months (rather than a year or more)

    …and the output is of higher quality than CASL's early dictionaries.
- These tools help–not replace–human beings.
- Contact information:
  - Dr. Mike Maxwell (Technical Director for HLT): `mmaxwell@casl.umd.edu` 301-356-2639
  - Dr. David Zajic (Research Scientist): `dzajic@casl.umd.edu` 301-356-8995
  - Dr. Mike Bunting (CASL Executive Director): `mbunting@casl.umd.edu` 301-356-8894

Current DYMs

- Arabic (Iraqi, MSA, Sudanese, Levantine...; dictionaries, gazetteers, 10,001 Arabic names); *French in Arabic (code switching)
- Urdu, Pashto, *Waziri Pashto, Punjabi
- Russian, Ukrainian
- Somali, Maay (related to Somali), Swahili, Chimwiini (related to Swahili)
- Persian (Farsi, Dari)
- Korean
- Mandarin Chinese (pinyin search)
- Dhivehi

*Not available in curent Platforms