# TextCritical.net - Task #479

Feature # 466 (Closed): Lemma lookup

## Diogenes Analyses Importer

12/17/2012 05:57 AM - Luke Murphey

| | | | |
|---|---|---|---|
| **Status:** | Closed | **Start date:** | 12/24/2012 |
| **Priority:** | Normal | **Due date:** | |
| **Assignee:** | Luke Murphey | **% Done:** | 100% |
| **Category:** | | **Estimated time:** | 0.00 hour |
| **Target version:** | 0.2 | | |

**Description**

Need to write an importer that takes the analyses file and marks up the lemma entries with the appropriate meaning.

**Subtasks:**

| | |
|---|---|
| Bug # 492: Analysis entries are not matching the extraction regular expression | **Closed** |

### History

**#1 - 12/17/2012 05:58 AM - Luke Murphey**

*- Status changed from New to In Progress*

**#2 - 12/17/2012 06:21 AM - Luke Murphey**

hmm, lining up the analyses file with the lemma is proving difficult. Consider the following which shows that matches are not occurring:

```
Forms discovered, count=2, line_number=100000, form_number=1
Forms discovered, count=2, line_number=100000, form_number=2
Forms discovered, count=2, line_number=100000, form_number=3
Forms discovered, count=2, line_number=100000, form_number=4
Forms discovered, count=1, line_number=100001, form_number=1
Forms discovered, count=1, line_number=100001, form_number=2
Forms discovered, count=1, line_number=100001, form_number=3
Forms discovered, count=1, line_number=100001, form_number=4
Forms discovered, count=1, line_number=100001, form_number=5
Forms discovered, count=1, line_number=100001, form_number=6
Forms discovered, count=1, line_number=100002, form_number=1
Forms discovered, count=1, line_number=100002, form_number=2
Forms discovered, count=1, line_number=100002, form_number=3
Forms discovered, count=1, line_number=100003, form_number=1
Forms discovered, count=1, line_number=100003, form_number=2
Forms discovered, count=1, line_number=100003, form_number=3
Forms discovered, count=1, line_number=100004, form_number=1
Forms discovered, count=1, line_number=100004, form_number=2
Forms discovered, count=1, line_number=100004, form_number=3
Forms discovered, count=0, line_number=100005, form_number=1
Forms discovered, count=0, line_number=100005, form_number=2
```

Here are lines 100000 through 100005:

```
a)nalou=to
    {7566000 9 a)na__lou=to,a)nali/skw    use up    imperf ind mp 3rd sg (doric aeolic)}
    {7566000 9 a)na_lou=to,a)nali/skw    use up    imperf ind mp 3rd sg (homeric ionic)}
    {7566000 9 a_)nalou=to,a)nalo/w    use up    imperf ind mp 3rd sg (doric aeolic)}
    {7566000 9 a)nalo/w    use up    imperf ind mp 3rd sg (homeric ionic)}
a)naloume/nas
    {7585355 9 a)naloume/na_s,a)na/llomai    leap    fut part mid fem acc pl (attic epic doric)}
    {7585355 9 a)naloume/na_s,a)na/llomai    leap    fut part mid fem gen sg (doric)}
    {7566000 9 a)na_loume/na_s,a)nali/skw    use up    pres part mp fem acc pl}
    {7566000 9 a)na_loume/na_s,a)nali/skw    use up    pres part mp fem gen sg (doric aeolic)}
    {7566000 9 a)naloume/na_s,a)nalo/w    use up    pres part mp fem acc pl}
    {7566000 9 a)naloume/na_s,a)nalo/w    use up    pres part mp fem gen sg (doric aeolic)}
a)naloume/nh
    {7585355 9 a)na/llomai    leap    fut part mid fem nom/voc sg (attic epic)}
    {7566000 9 a)na_loume/nh,a)nali/skw    use up    pres part mp fem nom/voc sg (attic epic ionic)}
    {7566000 9 a)nalo/w    use up    pres part mp fem nom/voc sg (attic epic ionic)}
a)naloume/nhn
    {7585355 9 a)na/llomai    leap    fut part mid fem acc sg (attic epic)}
    {7566000 9 a)na_loume/nhn,a)nali/skw    use up    pres part mp fem acc sg (attic epic ionic)}
    {7566000 9 a)nalo/w    use up    pres part mp fem acc sg (attic epic ionic)}
a)naloume/nhs
    {7585355 9 a)na/llomai    leap    fut part mid fem gen sg (attic epic)}
    {7566000 9 a)na_loume/nhs,a)nali/skw    use up    pres part mp fem gen sg (attic epic ionic)}
    {7566000 9 a)nalo/w    use up    pres part mp fem gen sg (attic epic ionic)}
a)naloume/nou
    {7585355 9 a)na/llomai    leap    fut part mid masc/neut gen sg (attic epic doric)}
    {7566000 9 a)na_loume/nou,a)nali/skw    use up    pres part mp masc/neut gen sg}
    {7566000 9 a)nalo/w    use up    pres part mp masc/neut gen sg}
```

**#3 - 12/17/2012 07:12 AM - Luke Murphey**

I wonder, should I just import the analyses file directly? Or rather, what does the lemmata file give me that the analyses file doesn't?

**#4 - 12/17/2012 07:59 AM - Luke Murphey**

From lemmata:

```
a(/bra    {537850 9 a(/bra_,a(/bra    favourite slave    fem nom/voc/acc dual}{537850 9 a(/bra_,a(/bra    favourite slave    fem nom/voc sg (attic doric aeolic)}
a(/brai    {537850 9 a(/bra    favourite slave    fem nom/voc pl}{537850 9 a(/bra_|,a(/bra    favourite slave    fem dat sg (attic doric aeolic)}
a(/brais    {537850 9 a(/bra    favourite slave    fem dat pl}
a(/bran    {537850 9 a(/bra_n,a(/bra    favourite slave    fem acc sg (attic doric aeolic)}
a(/bras    {537850 9 a(/bra_s,a(/bra    favourite slave    fem acc pl}{537850 9 a(/bra_s,a(/bra    favourite slave    fem gen sg (attic doric aeolic)}
a(/bra|    {537850 9 a(/brai,a(/bra    favourite slave    fem nom/voc pl}{537850 9 a(/bra_|,a(/bra    favourite slave    fem dat sg (attic doric aeolic)}
...
a(bra=n    {537850 9 a(/bra    favourite slave    fem gen pl (doric aeolic)}{555266 9 a(bro/s    graceful    masc/fem gen pl (doric)}
...
a(brw=n    {537850 9 a(/bra    favourite slave    fem gen pl}{555266 9 a(bro/s    graceful    fem gen pl}{555266 9 a(bro/s    graceful    masc/neut gen pl}{555266 9 a(bro/s    graceful    masc/fem/neut gen pl}
```

From the analysis:

```
a(/bra    537850    a(/bra (fem nom/voc/acc dual) (fem nom/voc sg (attic doric aeolic))    a(/brai (fem nom/voc pl) (fem dat sg (attic doric aeolic))    a(/brais (fem dat pl)    a(/bran (fem acc sg (attic doric aeolic))    a(/bras (fem acc pl) (fem gen sg (attic doric aeolic))    a(/bra| (fem nom/voc pl) (fem dat sg (attic doric aeolic))    a(bra=n (fem gen pl (doric aeolic))    a(brw=n (fem gen pl)
```

It seems like the lemma contains all forms of a lemma within a single line. The analysis file breaks up the forms onto each line with the line containing all possible meanings for the form. If this is the case, then I really need to be using the analyses file and not the lemma.

**#5 - 12/20/2012 04:56 AM - Luke Murphey**

The greek-analyses file has 911871 lines.

The following search will return the progress of the analyses import:

```
sourcetype="django" | stats max(line_number) as line_number | eval progress=100*line_number/911871
```

**#6 - 12/20/2012 04:59 AM - Luke Murphey**

You can clear out the entries in the database related to the Greek lemma:

```
drop table reader_lemma
drop table reader_case
drop table reader_dialect
drop table reader_worddescription
drop table reader_worddescription_cases
drop table reader_worddescription_dialects
drop table reader_wordform
```

**#7 - 12/20/2012 05:01 AM - Luke Murphey**

Some of the entries in the analyses file that are not in the greek-lemmata file. These can found with the following search:

```
sourcetype="django" "Unable to find the lemma for an analysis entry" | stats count(sourcetype)
```

**#8 - 12/20/2012 07:03 PM - Luke Murphey**

Successfully imported 911871 entries from the analyses file. However, the lemmas could not be found for 5,581 entries (see attached for a list).

The unmatched entries can be viewed with the following Splunk search:

```
sourcetype="django" "Unable to find the lemma for an analysis entry" | sort _time | table form line_number for
m_number
```

**#9 - 12/20/2012 07:21 PM - Luke Murphey**

*- File unmatched_analyses.csv added*

**#10 - 12/20/2012 08:01 PM - Luke Murphey**

It seems to take about six hours to import the analyses.

**#11 - 12/20/2012 08:01 PM - Luke Murphey**

*- % Done changed from 0 to 60*

**#12 - 12/20/2012 08:05 PM - Luke Murphey**

Some leftover issues:

1. ζῳάγρια lists the definition as "ransom paid for a prisoner taken alive) reward for life saved" (note the unmatched parenthesis)
2. Some entries are getting imported correctly because they have exclamation marks which are not valid beta-code from what I can tell so far

**#13 - 12/20/2012 08:24 PM - Luke Murphey**

*- % Done changed from 60 to 70*

**#14 - 12/21/2012 03:56 PM - Luke Murphey**

*- Tracker changed from Bug to Task*

**#15 - 12/22/2012 07:03 PM - Luke Murphey**

*- % Done changed from 70 to 90*

Some entries have leading characters that are preventing them from being matched accordingly.

**#16 - 12/24/2012 06:00 PM - Luke Murphey**

Lots of forms are being skipped. You can view them with the following search:

```
sourcetype="django" "Unable to find the lemma for an analysis entry" | table form line_number form_number
```

Also, many are not matching the regex. You can see these with the following search:

```
sourcetype="django" "Analysis entry does not match the regex" | table form line_number form_number
```

**#17 - 12/24/2012 06:14 PM - Luke Murphey**

The importer is now taking about 80 seconds to import 1,000 entries. The import started at 04:13:27 and has imported 352,000 at 2012-12-24 12:05:36. It is importing about 12 entries per second.

**#18 - 12/24/2012 06:25 PM - Luke Murphey**

Moving the commit to the file level did not improve performance noticeably.

**#19 - 12/24/2012 06:33 PM - Luke Murphey**

I tried deferring loading of all fields in the lemma query to speed up performance ( http://stackoverflow.com/questions/2846029/django-set-foreign-key-using-integer) but this didn't seem to help much.

**#20 - 12/24/2012 06:43 PM - Luke Murphey**

Time to load each entry seems to take about 0.03 seconds.

#### #21 - 12/24/2012 06:55 PM - Luke Murphey

The slowdown seems to be in the query of the lemmas for the lemma ID associated with the reference number. This is somewhat surprising since the reference field is indexed.

#### #22 - 12/24/2012 07:04 PM - Luke Murphey

Actually, the reference number field was not indexed for some reason. I manually created it with:

```
Create index "lemma_reference_number" ON "reader_lemma" ("reference_number");
```

#### #23 - 12/25/2012 04:22 AM - Luke Murphey

Many entries have single quotes at odd places ('ναι, εῦσθ'). Perhaps I should drop these entirely.

#### #24 - 12/27/2012 04:57 AM - Luke Murphey

*- Status changed from In Progress to Closed*

## Files

| | | | |
|---|---|---|---|
| unmatched_analyses.csv | 164 KB | 12/20/2012 | Luke Murphey |