TextCritical.net - Task #1229

Determine why searching for a particular division with chapter doesn't work

03/03/2016 12:17 AM - Luke Murphey

Status:	Closed	**Start date:**	03/03/2016
Priority:	Normal	**Due date:**	
Assignee:	Luke Murphey	**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:	1.3		

Description

The following should return nothing:

```
work:new-testament section:"Galatians 1" νόμον
```

Oddly enough, the following matches nothing:

```
work:new-testament section:"Galatians 10" νόμον
```

This does seem to work in other places. For example, the following works, returning only one result:

```
work:new-testament section:"Romans 13" νόμον
```

But this return too much:

```
work:new-testament section:"Romans 1" νόμον
```

History

#1 - 03/03/2016 12:18 AM - Luke Murphey

It seems like section filtering only works correctly when the chapter is 10 or greater.

#2 - 03/03/2016 12:19 AM - Luke Murphey

- Description updated

#3 - 03/03/2016 04:37 AM - Luke Murphey

Something happening in the query parser.

Original:
work:new-testament section:"Acts 1""

Parsed:
(work:<new> AND work:<testament> AND section:acts)

Original:
work:new-testament section:"Acts 10"

Parsed:
(work:<new> AND work:<testament> AND section:"acts 10")

#4 - 03/03/2016 04:46 AM - Luke Murphey

Observations:

I don't like the that the following is happening:

1. That the single digit numbers are being dropped in the section
2. That the dash is being used to split up the work name

#5 - 03/03/2016 04:54 AM - Luke Murphey

```
from whoosh.analysis import SimpleAnalyzer
ana = SimpleAnalyzer()
[token.text for token in ana(u"new-testament")]
```

Outputs:
[u'new', u'testament']

#6 - 03/03/2016 04:57 AM - Luke Murphey

I think the problem is that Whoosh isn't recognizing that the section is quoted and thus is hitting the minsize limit:
http://whoosh.readthedocs.org/en/latest/api/analysis.html?highlight=SimpleAnalyzer

#7 - 03/03/2016 05:11 AM - Luke Murphey

- Status changed from New to In Progress

#8 - 03/03/2016 06:58 AM - Luke Murphey

This returns "Acts 1" as expected:

```
from whoosh.analysis import SimpleAnalyzer
from whoosh.util import rcompile
ana = SimpleAnalyzer( rcompile(r"[a-zA-Z0-9- ]+") )
[token.text for token in ana(u"section:acts 1")]
```

#9 - 03/03/2016 07:15 AM - Luke Murphey

I tried with both of the following and they seem to work:

- section_analyzer = StandardAnalyzer(rcompile(r"[a-zA-Z0-9-]+"), minsize=1)
- section_analyzer = SimpleAnalyzer(rcompile(r"[a-zA-Z0-9-]+"))

#10 - 03/03/2016 07:16 AM - Luke Murphey

There is one issue. I have lost the ability to search for works without a full section. In other words, this no longer works:

```
work:new-testament section:"Galatians" νόμον
```

#11 - 03/03/2016 07:46 AM - Luke Murphey

Going to try building the search indexes with more ways to refer to the divisions.

#12 - 03/03/2016 08:07 AM - Luke Murphey

The issue is that I am only able to now search for the first chapter description in get_section_index_text()'s output.

#13 - 03/03/2016 08:29 AM - Luke Murphey

- Status changed from In Progress to Closed

- % Done changed from 0 to 100