INL LEXICONSERVICE - MANUAL

Author: Mathieu Fannee, INL – March 2016

CONTENT

INTRODUCTION

The INL LexiconService is a webservice that gives any piece of software quick online access to a lexicon by means of http requests. The LexiconService is designed to access a computational lexicon with an Impact Lexicon Database structure.[1] The service is now deployed at INL and gives access to INL's GiGaNT lexicon of 13th to 20th century Dutch.

This webservice offers various possibilities. One can obtain the word forms belonging to a given lemma, or the other way round, one can get the lemma corresponding to a given word form. It is also possible to expand any word with its complete paradigm. And one can limit the results to a given period of history, or to a given part-of-speech. Finally, the lexical information provided by the webservice can be given in both XML or JSON format.

In the following, we'll be describing how requests need to be formulated to get the information you need.

QUERY BASICS

To be able to get lexical information from the LexiconService, you need to provide it with at least three things:

- A word
- A lexicon to look up this word in
- And what you need to get in return (a lemma, a paradigm, …)

Let's start with the lexicon in which the word should be looked up. Let's say we want to access some Dutch lexicon, which is named '*lexicon_service_db*'.[2] Your request will have to contains this part:

```
...database=lexicon_service_db...
```

Then you have to tell the LexiconService what you need to get. Three distinct operations are possible:

- Get the lemma of a word form
- Get the word forms of a lemma (=its paradigm)
- Expand a word to its complete paradigm (=lemma and all word forms)

Telling the LexiconService which word your question is about can be done in different ways, depending on the operation you need to be performed. So we're going to describe these three possible operations now.

[1] The current Dutch lexicon consists of a lemmata table with grammatical categories (part of speech), a wordforms table, an attestations table, a documents table with metadata (source information and date).
[2] The list of available lexica depends on the specific LexiconService instance you're using and also on the period of time (as new lexica might be added and old ones removed). Please contact the administrator of the LexiconService instance you are using to get an up-to-date list of the available lexica.

GET A LEMMA

Telling the LexiconService to get the lemma of a word form is done by simply telling 'get_lemma'. The first part of your http requests will therefore look like this:

```
http://lexiconservice.inl.nl/LexiconService/lexicon/get_lemma?...
```

BEWARE: *lexiconservice.inl.nl* (mentioned here above) is just a sample hostname, so this host might not be available anymore as you get to read this. Check the document about the currently available datasets (or contact the INL) to get the URL at which the service can currently be reached.

You then need to set a few parameters: the word form you want the lemma from, and the name of the lexicon where to look it up. Let's say your word form is *liep* (the Dutch for *walked*). Your complete request will look like this:

```
http://lexiconservice.inl.nl/LexiconService/lexicon/get_lemma?
database=lexicon_service_db&wordform=liep
```

Provided that you want JSON output, the LexiconService will send a response like this:

```
{
"message":"OK",
"lemmata_list":
    [{"query_word":"liep",
     "found_lemmata":
         [{"lemma":"lijp","pos":"ADJ ADV"},
          {"lemma":"lopen","pos":"VRB"}]
    }]
}
```

GET WORDFORMS

Telling the LexiconService to get the word forms of a lemma is done by simply telling 'get_wordforms'. The first part of your http requests will therefore look like this:

```
http://lexiconservice.inl.nl/LexiconService/lexicon/get_wordforms?...
```

You then need to set a few parameters: the lemma you want the word forms from, and the name of the lexicon where to look it up. Let's say your lemma is *lopen* (the Dutch for *to walk*). Your complete request will look like this:

```
http://lexiconservice.inl.nl/LexiconService/lexicon/get_wordforms?
database=lexicon_service_db&lemma=lopen
```

Provided that you want JSON output, the LexiconService will send a response like this:

```
{
"message":"OK",
"wordforms_list":
    [{"query_word":"lopen",
     "found_wordforms":["loop","loope","lopen", …]
    }]
}
```

EXPAND A WORDFORM OR LEMMA

Telling the LexiconService to expand a given word form or lemma to its complete paradigm is done by simply telling 'expand'. The first part of your http requests will therefore look like this:

```
http://lexiconservice.inl.nl/LexiconService/lexicon/expand?...
```

You then need to set a few parameters: the lemma you want the word forms from, and the name of the lexicon where to look it up. Let's say your word is *loopt* (the Dutch for *(he) walks*). Your complete request will look like this:

```
http://lexiconservice.inl.nl/LexiconService/lexicon/expand?
database=lexicon_service_db&wordform=loopt
```

Provided that you want JSON output, the LexiconService will send a response like this:


```
{
"message":"OK",
"wordforms_list":
    [{"query_word":"loopt",
      "found_wordforms":["eloopen","geloopen","gelopen","gheloopen",
        "ghelooppen","laupe","liep","liepen","liept","loepen","loop","loope",
        "loopen","loopend","loopende","loopenden","loopens","looppen","looppene",
        "loopt","lopen","lopende"]
    }]
}
```


BEWARE: the LexiconService does not only expand a word with its complete paradigm but also with its lemma. In some cases, this behaviour might be inappropriate, for example because the paradigm consists of historical words whereas the lemma is a modern form. In such cases, you might want to exclude the lemma from the expansion results. To do so, just add the 'exclude_lemma' parameter to the request:

```
http://lexiconservice.inl.nl/LexiconService/lexicon/expand?
database=lexicon_service_db&wordform=loopt& exclude_lemma=true
```

ADVANCED QUERIES

LIMIT THE SEARCH TO SOME PART-OF-SPEECH

We've seen before how to get the lemma of a given word form. The needed query for the Dutch word form *liep* was:

```
http://lexiconservice.inl.nl/LexiconService/lexicon/get_lemma?
database=lexicon_service_db&wordform=liep
```

The result consisted of two very different lemmata, a verb and an adjective/adverb:


```
{
"message":"OK",
"lemmata_list":
    [{"query_word":"liep",
     "found_lemmata":
         [{"lemma":"lijp","pos":"ADJ ADV"},
          {"lemma":"lopen","pos":"VRB"}]
    }]
}
```

Now imagine you're not interested in adjectives (ADJ), but only in verbs (VRB).[3] You can set an extra parameter 'pos' (part-of-speech) to say just that:

```
http://lexiconservice.inl.nl/LexiconService/lexicon/get_lemma?
database=lexicon_service_db&wordform=liep&pos=VRB
```

Now the LexiconService output will be:

```
{
"message":"OK",
"lemmata_list":
    [{"query_word":"liep",
      "query_pos":"VRB",
      "found_lemmata":
          [{"lemma":"lopen","pos":"VRB"}]
    }]
}
```

Exactly the same can be achieved for the other operations 'get_wordforms' and 'expand'.

The request for word forms of lemma *lopen* limited to a part-of-speech *VRB* will look like:

```
http://lexiconservice.inl.nl/LexiconService/lexicon/get_wordforms?
database=lexicon_service_db&lemma=lopen&pos=VRB
```

And the request for expansion of *loopt* limited to a part-of-speech *VRB* will be:

```
http://lexiconservice.inl.nl/LexiconService/lexicon/expand?
database=lexicon_service_db&wordform=loopt&pos=VRB
```

LIMIT THE SEARCH TO A PERIOD OF TIME

The LexiconService offers the possibility to limit a search to a given period of time. This can be achieved by adding two parameters, 'year_from' and 'year_to'. It is possible to use only one of them, of both at the same time.

Let's say you'd like to get word forms of the word *lopen* before the year 1700, your request will look like:

```
http://lexiconservice.inl.nl/LexiconService/lexicon/get_wordforms?
database=lexicon_service_db&lemma=lopen&year_to=1700
```

This will give quite some oldish word forms:

```
… "found_wordforms":[ "gheloopen","ghelooppen",… ]} …
```

To get the modern word forms (after 1900) of *lopen* instead, we can use 'year_from':

```
http://lexiconservice.inl.nl/LexiconService/lexicon/get_wordforms?
database=lexicon_service_db&lemma=lopen&year_from=1900
```

With more modern forms as a result:

```
… "found_wordforms":[ "geloopen","liep","loopen","loopend","loopende","loopt" ]} …
```

[3] The parts-of-speech in use in a given lexicon depends on the choices made by the authors of the lexicon in question. Please contact the authors of the lexicon you'd like to use to get a list of relevant parts-of-speech.

Of course, the ‘year_from’ and ‘year_to’ parameters can be used together so as to isolate a given period of time. Say we want the paradigm of *lopen* in the period 1600-1700, our request will be:

```
http://lexiconservice.inl.nl/LexiconService/lexicon/get_wordforms?
database=lexicon_service_db&lemma=lopen&year_from=1600&year_to=1700
```

QUERYING MORE WORDS AT ONCE

In the sections hereabove we explored the possibilities for getting paradigm information and such for one single word only. But the LexiconService can process lists of words as well.

Sending a query for list of words is easy, just separate the different words by comma's:

```
http://lexiconservice.inl.nl/LexiconService/lexicon/get_lemma?
database=lexicon_service_db&wordform=liep,werk,dacht
```

The result will be expectably:

```
{
"message":"OK",
"lemmata_list":
    [{"query_word":"liep",
      "query_pos":"",
      "found_lemmata":
          [{"lemma":"lopen","pos":"VRB"}]
     },
     {"query_word":"werk",
      "query_pos":"",
      "found_lemmata":
          [{"lemma":"werk","pos":"NOU"}]
     },
     {"query_word":"dacht",
      "query_pos":"",
      "found_lemmata":
          [{"lemma":"denken","pos":"VRB"}]
     }]
}
```

As we saw before, when writing a query about a word, it is possible to specify its part of speech. When dealing with lists of words, the parts of speech need to be comma separated as well.

So, querying about the words *liep* ('walked'), *man* ('husband'), *aardig* ('nice', 'sweet'), the parts of speech information being respectively VRB (verb), NOU (noun) and ADJ (adjective), the resulting query will be:

```
http://lexiconservice.inl.nl/LexiconService/lexicon/get_lemma?
database=lexicon_service_db&wordform=liep,man,aardig&pos=VRB,NOU,ADJ
```

Make sure, when adding parts of speech information, that each word form or lemma is provided an own part of speech. That is: the number of word forms or lemmata must equal the number of parts of speech. Otherwise the LexiconService won't be able to match the lemmata with the parts of speech information, and you will get an error.

CASE SENSITIVITY

By default, the LexiconService is case sensitive. So querying the word *liep* or the word *Liep* (same word, but front letter in uppercase) might give different results.

If you want the LexiconService to be case *insensitive* instead, we can require just that by adding one parameter: 'case_sensitive=true/false'. Like that:

```
http://lexiconservice.inl.nl/LexiconService/lexicon/get_lemma?
database=lexicon_service_db&wordform=liep&case_sensitive=false
```

ERROR INFORMATION

When the LexiconService needs to send out an error message, it puts this message into the JSON of XML response.

For example:

```
{
 "message":"ERROR: here comes your error message.",
 "lemmata_list":[]
}
```

The error messages of the LexiconService are designed to be as self-explanatory as possible. Error messages clearly tell which part of the input is missing or illegal, so finding the right way to solve a problem is mostly straightforward.
If not, don't hesitate to contact the administrator of the LexiconService instance you are using.

PREVENT CACHING

As part of their optimization strategies, some servers might cache requests and responses, in such a way that they can reply faster and with less CPU use when receiving a request they had to process before. A bad thing about this is that if you're working with a growing lexicon, you won't be able to get newly added information for a word you already send a request about.

This can be solved easily, just by adding a 'dummy' parameter with some random number to the http request, in such a way that the request will always look different from requests sent before, even if those were about the same word. For example:

```
http://lexiconservice.inl.nl/LexiconService/lexicon/get_lemma?
database=lexicon_service_db&wordform=liep&dummy=1384187319550
```

OUTPUT TYPE

The LexiconService can give both XML and JSON output. The output type cannot be set by an explicit parameter of the http request: you have to set it in the AJAX call of the application you're using to connect to the LexiconService.

UPGRADE FROM PREVIOUS VERSION

If you've been using an old version of the LexiconService (that is: **installed before august 2014**), you'll notice the output of the service has changed a bit. So, if you've implemented the LexiconService within your own software, you'll need to slightly adapt the part of your code which parses the LexiconService output.

In the old version, the output consisted of a *lemma* object or a *wordform* object. Those objects could only contain one set of results, about only one query word. Since the new versions of the LexiconService can process multiple query words at once (all sent within one single request, see section '*Querying more words at once*'), the output now contains a list of results instead of a single result: one set per query word.
So, the *lemma* and *wordform* objects are now embedded within a *lemmata_list* or a *wordforms_list* object respectively. Those lists contain the original *lemma* and *wordform* objects, which are now renamed *found_lemmata* and *found_wordforms*. They were renamed that way to distinguish them from the queried wordforms, which are stored in *query word*.

In the following table, the output differences between the old version (till august 2014) and the current version of the LexiconService are summed up, so you can see in a glance what has changed.

Task	Output of old version (before august 2014)	Output of current version

Get lemma		`{`
		`"message":"OK",`
		`"lemmata_list":[`

	`{`
	`"lemmata":[`
	`{"lemma":"lopen","pos":"VRB"},`
	`{"lemma":"lijp","pos":"ADJ"}`
	`]`
	`}`
		`{`
		`"query_word":"liep",`
		`"found_lemmata":[`
		`{"lemma":"lopen","pos":"VRB"},`
		`{"lemma":"lijp","pos":"ADJ"}`
		`]`
		`},`
		`{`
		`"query_word":"hoorde",`
		`...`
		`}`

		`]}`

Get wordforms		`{`
		`"message":"OK",`
		`"wordforms_list":[`

	`{`
	`"wordform":[`
	`"loop","loope","lopen", …`
	`]`
	`}`
		`{`
		`"query_word":"lopen",`
		`"found_wordforms":[`
		`"loop","loope","lopen", …`
		`]`
		`},`
		`{`
		`"query_word":"horen",`
		`...`
		`}`

		`]}`

Expand a word of lemma		`{`
		`"message":"OK",`
		`"wordforms_list":[`

	`{`
	`"wordform":[`
	`"geloopen","gelopen", …`
	`]`
	`}`
		`{`
		`"query_word":"loopt",`
		`"found_wordforms":[`
		`"geloopen","gelopen", …`
		`]`
		`},`
		`{`
		`"query_word":"hoorde",`
		`...`
		`}`

		`]}`