Collaborative notebook

A space for collaborative notetaking (if you want!).

Instructions

- Keep this document tidy 😊

Useful links

- LiLa LP6.2 activity website: https://lila-erc.eu/linked-pasts-6-activity/
- Linked Pasts website: https://ics.sas.ac.uk/events/linked-pasts-6
- Linked Pasts 6: activity details and links: https://docs.google.com/document/d/15OYNtGwImKfnkP-7dyTC1qRLLcBaFprZTI-OVoh ePCs/edit#
- LiLa Lemma Bank query interface: https://lila-erc.eu/query/
- LiLa SPARQL interface: https://lila-erc.eu/sparql/
- **LiLa TextLinker: http://lila-erc.eu:8080/LiLaTextLinker/**
- LiLa LodLive interface: https://lila-erc.eu/lodlive/

Activity workflow

Objective: connect a Latin text to the LiLa Knowledge Base of Linguistic Resources.

Navigate to the LiLa Text Linker at: http://lila-erc.eu:8080/LiLaTextLinker/

Copy and paste a Latin text (in plain text format, TXT) in the box provided and click on the PROCESS button in top-right corner of your screen. You'll now see that your text is highlighted in three different colours. Why? The Text Linker has mapped your text to the LiLa Lemma Bank and highlighted each token (word) in the text in one of three colours: green indicates an exact match, blue an ambiguous match and red a no match. What the Text Linker has also done is parse the text for part of speech (PoS) and lemma. If you hover over a word in your text, a pop-up window will show you the PoS and lemma of that word. If you click on that same word, a pane will open on the right hand side of your screen with all the information associated with that lemma in the LiLa Knowledge Base: PoS, lemma, lemma ID, etc.

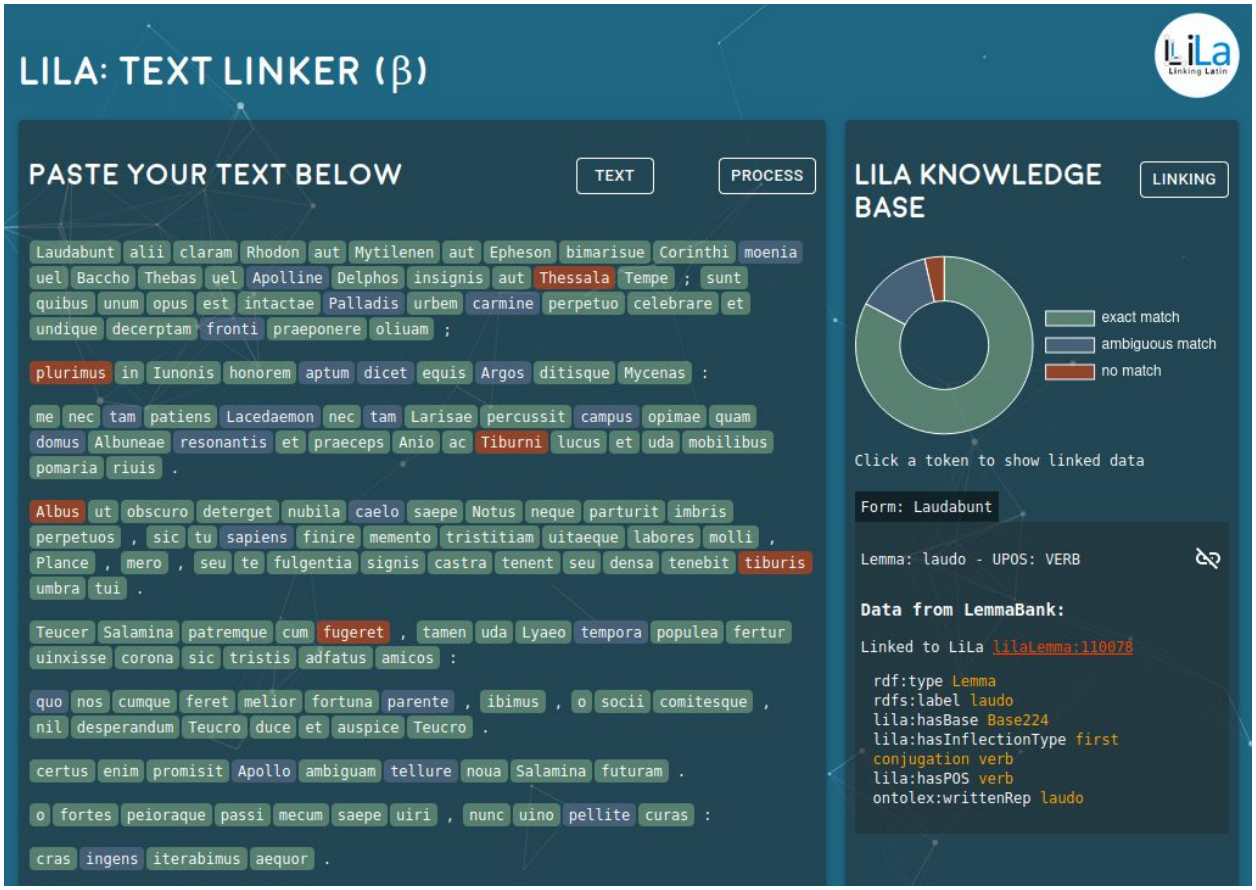


Fig. 1: Processed output of the LiLa Text Linker. Clicking on a word in the text, e.g. *laudabunt*, wil open a pane on the right-hand side of the screen displaying all of the information associated with that word in the LiLa Knowledge Base.

Clicking on the lemma ID in this pane (e.g. `lilaLemma:118878` in Fig. 1), will redirect you to the lemma's descriptive sheet (Fig. 2).

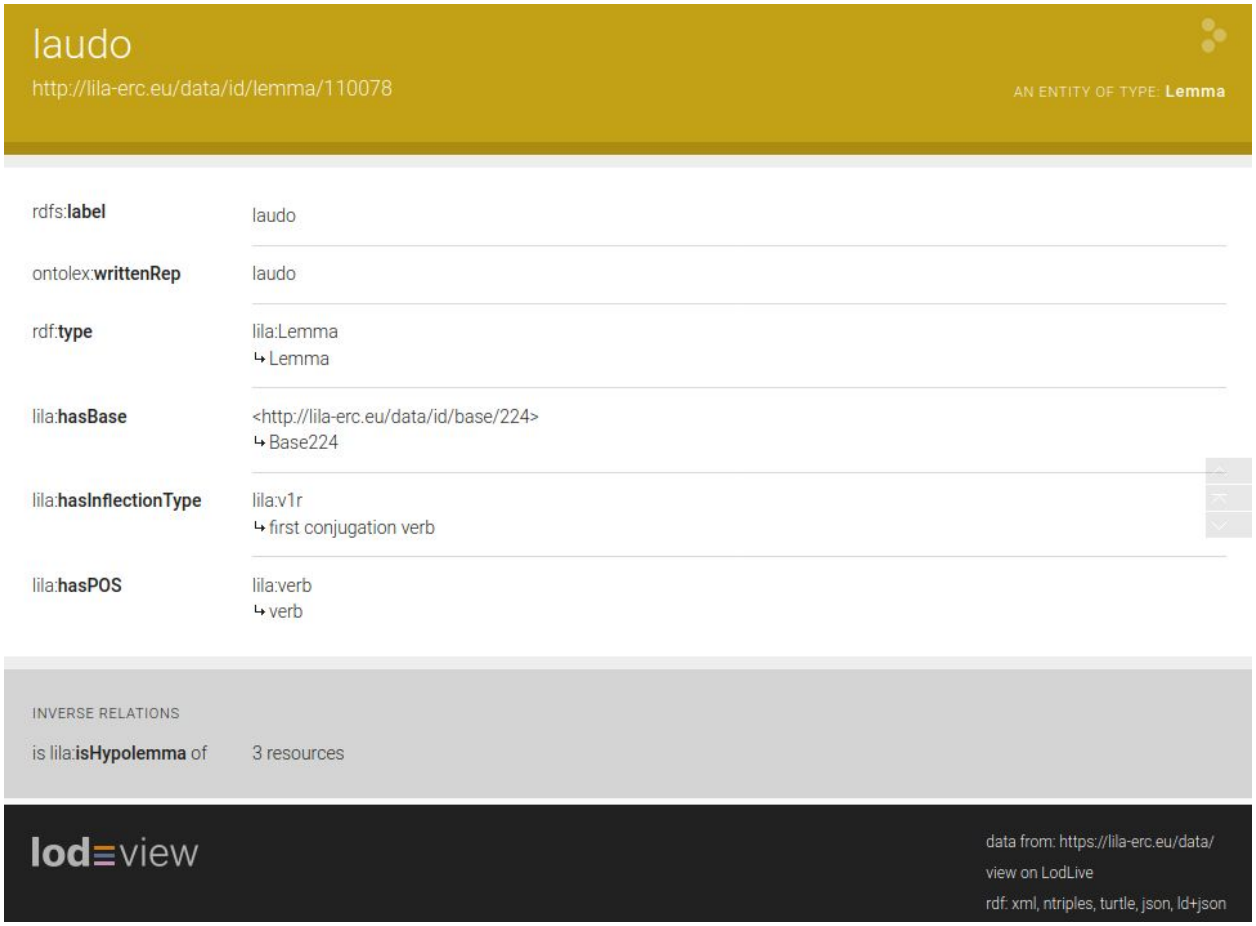


Fig. 2: Descriptive sheet of the lemma *laudo* in LiLa.

If, in this sheet, you click on the three-dot icon in the top-right corner of your screen will redirect you to the graph view of *laudo* (Fig. 3). Clicking on the node's "satellite" nodes will open up all connections associated with that lemma in LiLa!

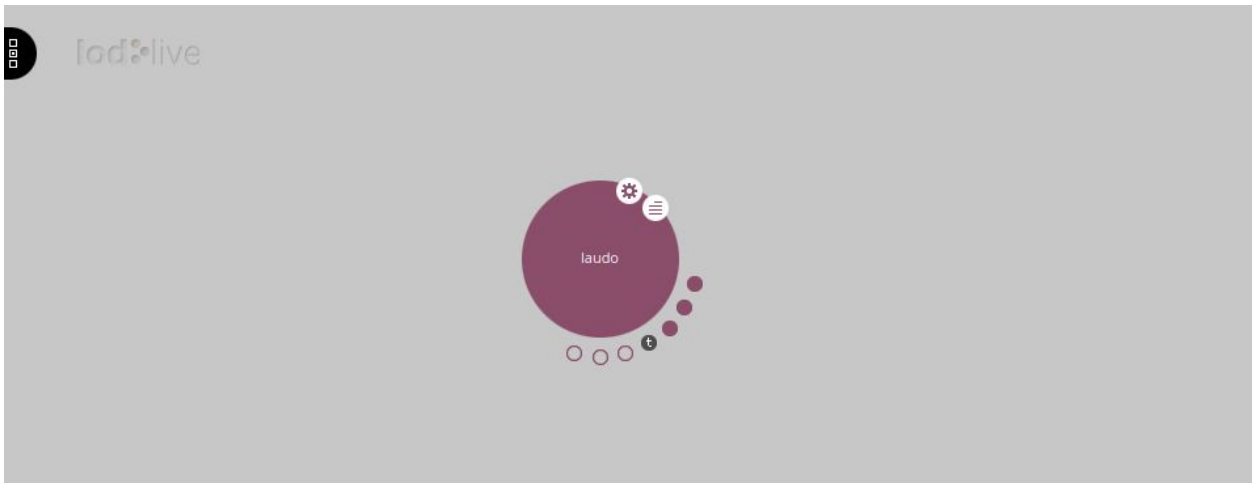


Fig. 3: Graph representation of *laudo* in LiLa.

Now, back to the Text Linker! Let us click on the red-highlighted lemma *plurimus* (Fig. 4). Doing so will open up the information pane on the right-hand side of the screen, which, as you'll see, is empty because this word has not been successfully matched against a lemma in LiLa.

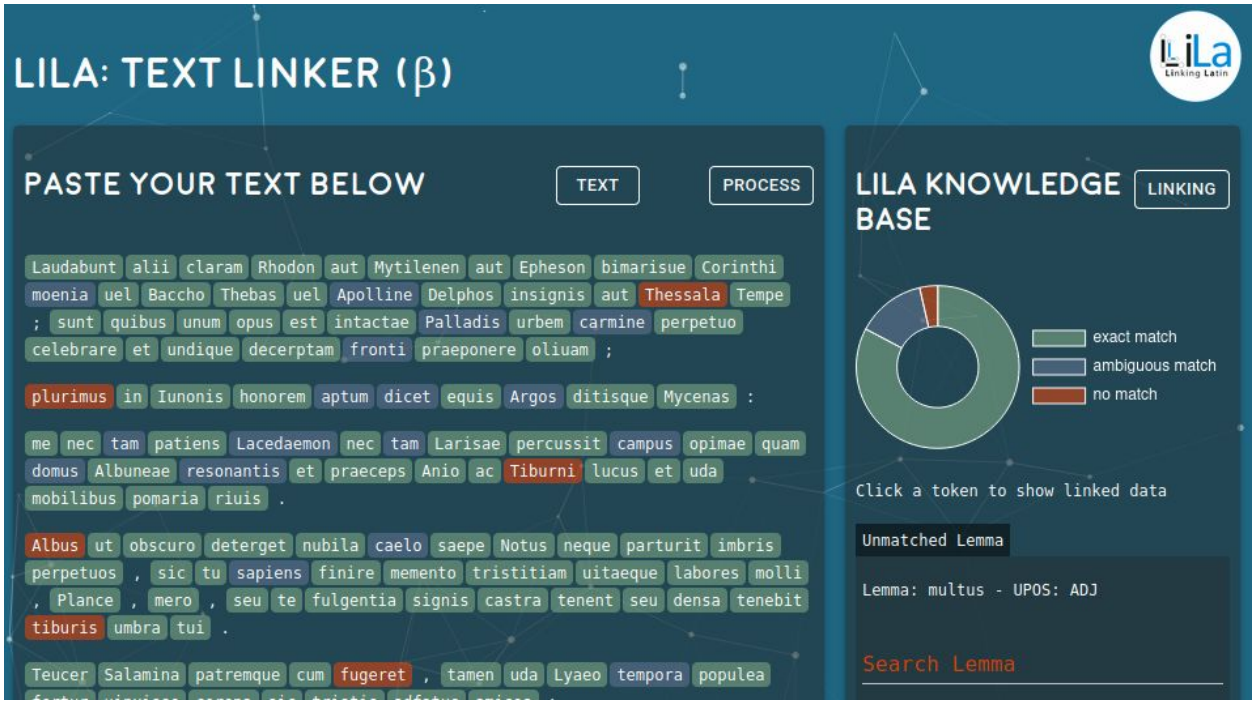


Fig. 4: Fixing a red-highlighted word (a "no match").

We need to create this association! To do so, using the "Search Lemma" functionality in that pane, we start typing the lemma of *plurimus*, that is, *multus* (Fig. 5). As you type, the Text Linker will provide suggestions. In our case, the only suggestion provided is `multus - determiner`.

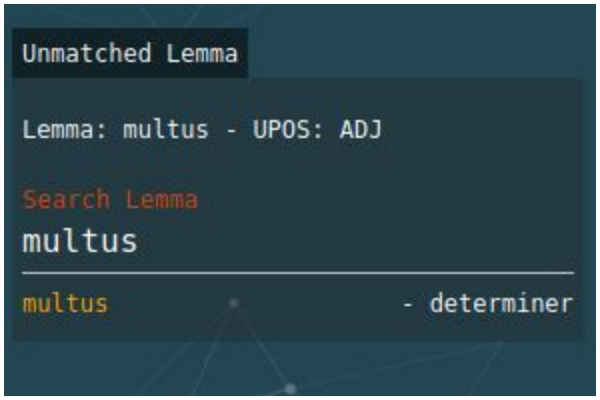


Fig. 5: Adding a LiLa lemma association to *plurimus*.

Clicking on that suggestion will automatically link our red *plurimus* to the lemma *multus*! What you should see now is a screen with all of the information associated with the lemma *multus* in LiLa (Fig. 6).

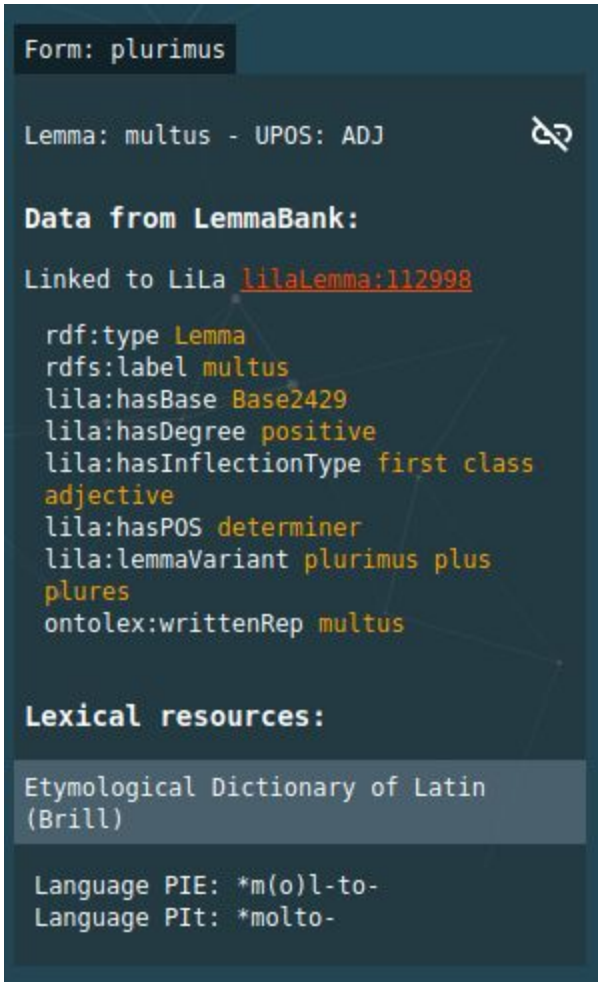


Fig. 6: The non-matched lemma *plurimus* has been successfully connected to the LiLa lemma *multus*.

As you'll now notice, this operation has changed the colour of *plurimus* in your text from red to green.

Let's now turn to ambiguously-matched lemmas, those highlighted in blue in the TextLinker (Fig. 7). If you click on an ambiguous lemma, the TextLinker will provide you with possibilities for you to choose from. Look for the correct lemma - when you find it click on the CHOOSE button to select it. Choosing a lemma will change the colour of that word from blue to green.

Once you've cleaned the text of red- and blue-highlighted words, you may proceed to linking it to the LiLa Knowledge Base. "Linking" here means converting our clean, lemmatised text to the Turtle syntax (TTL, in short) required by LiLa to publish your text in the Knowledge Base.

To link the text, click on the LINKING button in the top-right corner of your Text Linker screen. A Dialog box or form will appear asking you to fill-in some details about your text (Fig. 7). Complete the form and then press on GENERATE TTL to generate and download the Turtle version of your text.

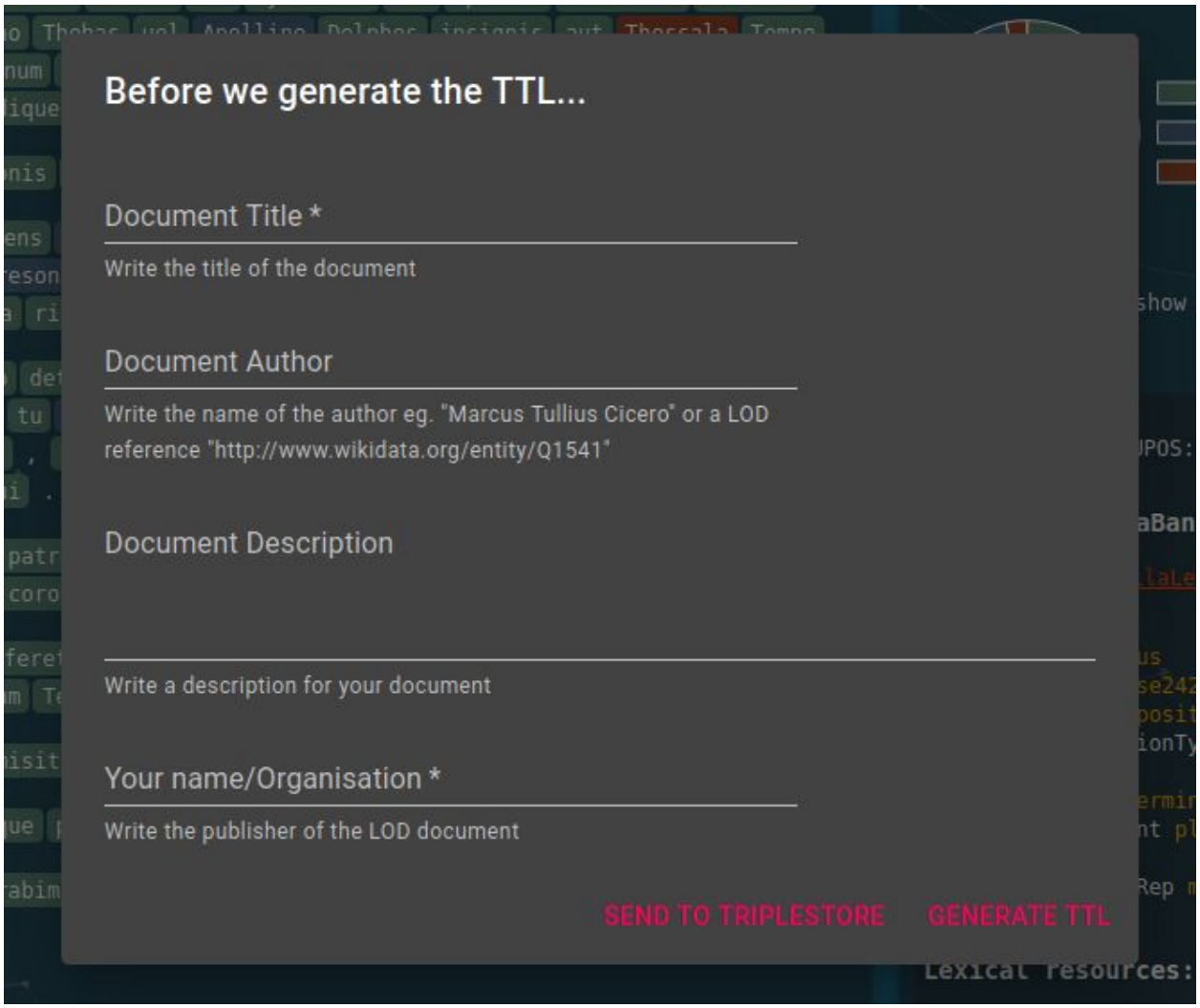
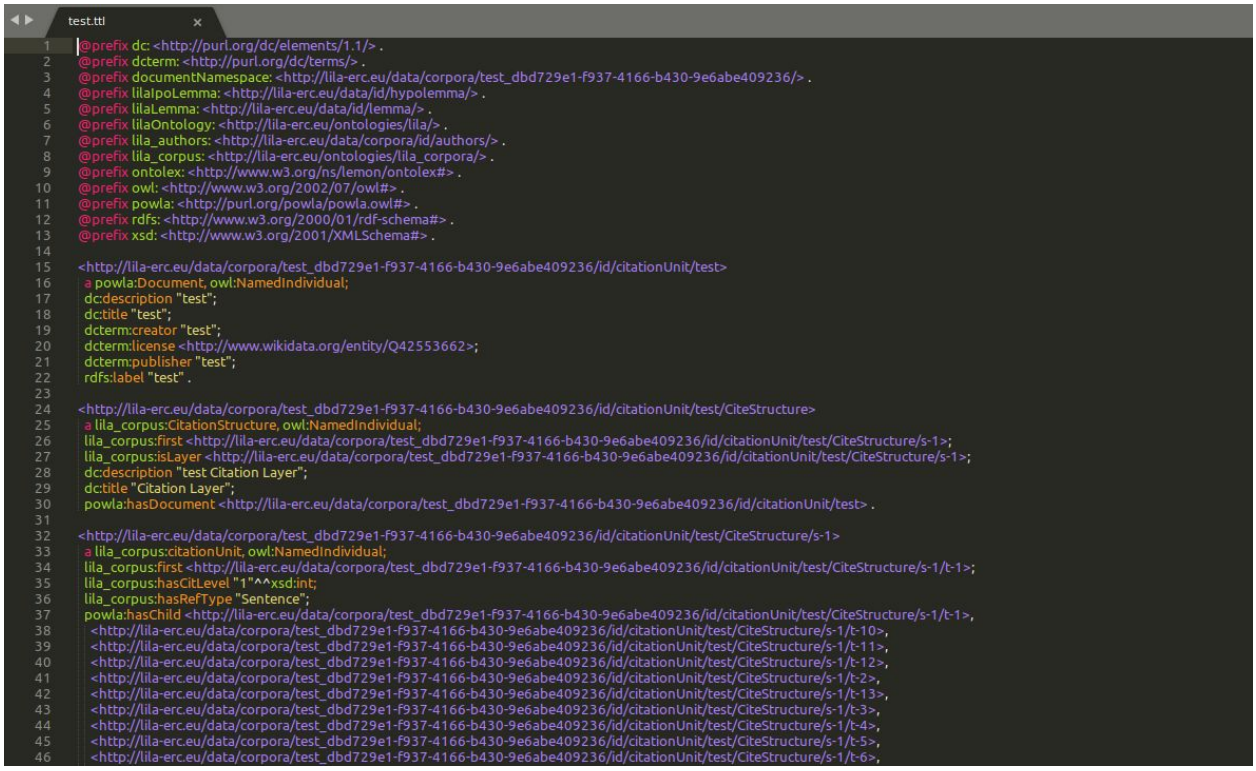


Fig. 7: Dialog box to generate and download a TTL version of your clean, lemmatised Latin text.

You can now open the downloaded `.ttl` file on your computer in any text editor (e.g. Notepad, Sublime Text Editor, Atom, etc.). This is what that file looks like (Fig. 8):



```
test.ttl
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix dcterm: <http://purl.org/dc/terms/> .
@prefix documentNamespace: <http://lila-erc.eu/data/corpora/test_dbd729e1-f937-4166-b430-9e6abe409236/> .
@prefix lilaIpoLemma: <http://lila-erc.eu/data/id/hypolemma/> .
@prefix lilaLemma: <http://lila-erc.eu/data/id/lemma/> .
@prefix lilaOntology: <http://lila-erc.eu/ontologies/lila/> .
@prefix lila_authors: <http://lila-erc.eu/data/corpora/id/authors/> .
@prefix lila_corpus: <http://lila-erc.eu/ontologies/lila_corpora/> .
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix powla: <http://purl.org/powla/powla.owl#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<http://lila-erc.eu/data/corpora/test_dbd729e1-f937-4166-b430-9e6abe409236/id/citationUnit/test>
 a powla:Document, owl:NamedIndividual;
 dc:description "test";
 dc:title "test";
 dcterm:creator "test";
 dcterm:license <http://www.wikidata.org/entity/Q42553662>;
 dcterm:publisher "test";
 rdfs:label "test" .

<http://lila-erc.eu/data/corpora/test_dbd729e1-f937-4166-b430-9e6abe409236/id/citationUnit/test/CiteStructure>
 a lila_corpus:CitationStructure, owl:NamedIndividual;
 lila_corpus:first <http://lila-erc.eu/data/corpora/test_dbd729e1-f937-4166-b430-9e6abe409236/id/citationUnit/test/CiteStructure/s-1>;
 lila_corpus:isLayer <http://lila-erc.eu/data/corpora/test_dbd729e1-f937-4166-b430-9e6abe409236/id/citationUnit/test/CiteStructure/s-1>;
 dc:description "test Citation Layer";
 dc:title "Citation Layer";
 powla:hasDocument <http://lila-erc.eu/data/corpora/test_dbd729e1-f937-4166-b430-9e6abe409236/id/citationUnit/test> .

<http://lila-erc.eu/data/corpora/test_dbd729e1-f937-4166-b430-9e6abe409236/id/citationUnit/test/CiteStructure/s-1>
 a lila_corpus:citationUnit, owl:NamedIndividual;
 lila_corpus:first <http://lila-erc.eu/data/corpora/test_dbd729e1-f937-4166-b430-9e6abe409236/id/citationUnit/test/CiteStructure/s-1/t-1>;
 lila_corpus:hasCitLevel "1"^^xsd:int;
 lila_corpus:hasRefType "Sentence";
 powla:hasChild <http://lila-erc.eu/data/corpora/test_dbd729e1-f937-4166-b430-9e6abe409236/id/citationUnit/test/CiteStructure/s-1/t-1>,
  <http://lila-erc.eu/data/corpora/test_dbd729e1-f937-4166-b430-9e6abe409236/id/citationUnit/test/CiteStructure/s-1/t-10>,
  <http://lila-erc.eu/data/corpora/test_dbd729e1-f937-4166-b430-9e6abe409236/id/citationUnit/test/CiteStructure/s-1/t-11>,
  <http://lila-erc.eu/data/corpora/test_dbd729e1-f937-4166-b430-9e6abe409236/id/citationUnit/test/CiteStructure/s-1/t-12>,
  <http://lila-erc.eu/data/corpora/test_dbd729e1-f937-4166-b430-9e6abe409236/id/citationUnit/test/CiteStructure/s-1/t-2>,
  <http://lila-erc.eu/data/corpora/test_dbd729e1-f937-4166-b430-9e6abe409236/id/citationUnit/test/CiteStructure/s-1/t-13>,
  <http://lila-erc.eu/data/corpora/test_dbd729e1-f937-4166-b430-9e6abe409236/id/citationUnit/test/CiteStructure/s-1/t-3>,
  <http://lila-erc.eu/data/corpora/test_dbd729e1-f937-4166-b430-9e6abe409236/id/citationUnit/test/CiteStructure/s-1/t-4>,
  <http://lila-erc.eu/data/corpora/test_dbd729e1-f937-4166-b430-9e6abe409236/id/citationUnit/test/CiteStructure/s-1/t-5>,
  <http://lila-erc.eu/data/corpora/test_dbd729e1-f937-4166-b430-9e6abe409236/id/citationUnit/test/CiteStructure/s-1/t-6>,
```

Fig. 8: Version of our Latin text in the Turtle syntax.

Now, let us send our TTL to LiLa's Knowledge Base! In the previously shown Dialog box (Fig. 7), click on SEND TO TRIPLESTORE, copy the link provided by the system and put it somewhere safe (Fig. 9)

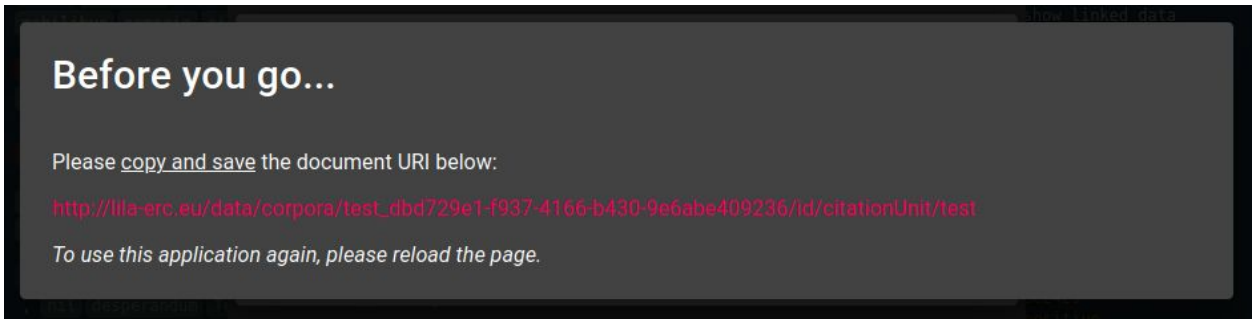


Fig. 9: LiLa URL of your Latin text.

If you paste that link in the URL field of your browser you will be redirected to the LiLa page of your text (Fig. 10).

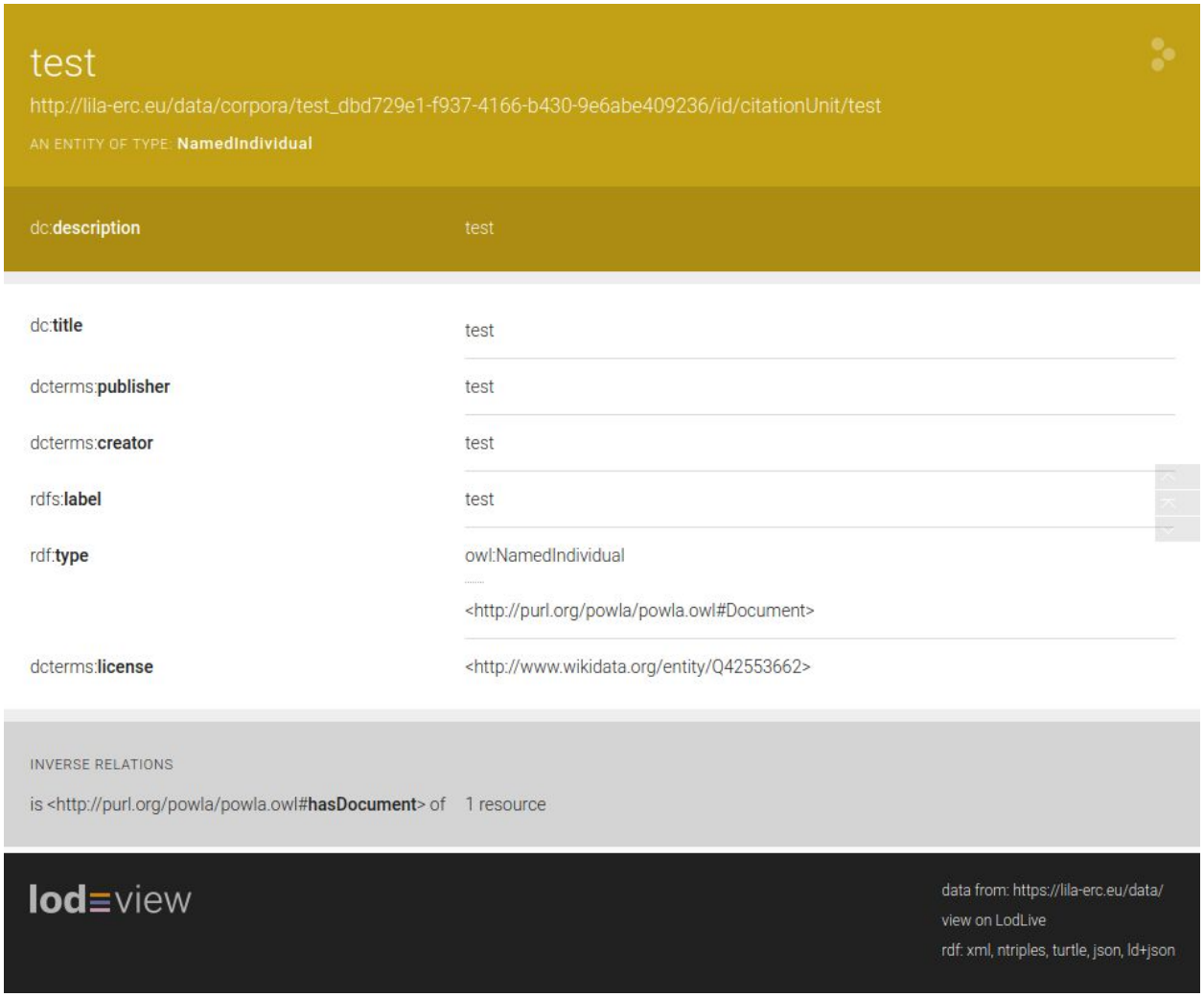


Fig. 10: LiLa view of our Latin text.

Now, clicking on the three-dot icon in the top-right corner of this screen will open up the graph representation of your text (Fig. 11)! You can now explore your text in relation to all other information already contained in the LiLa Knowledge Base.

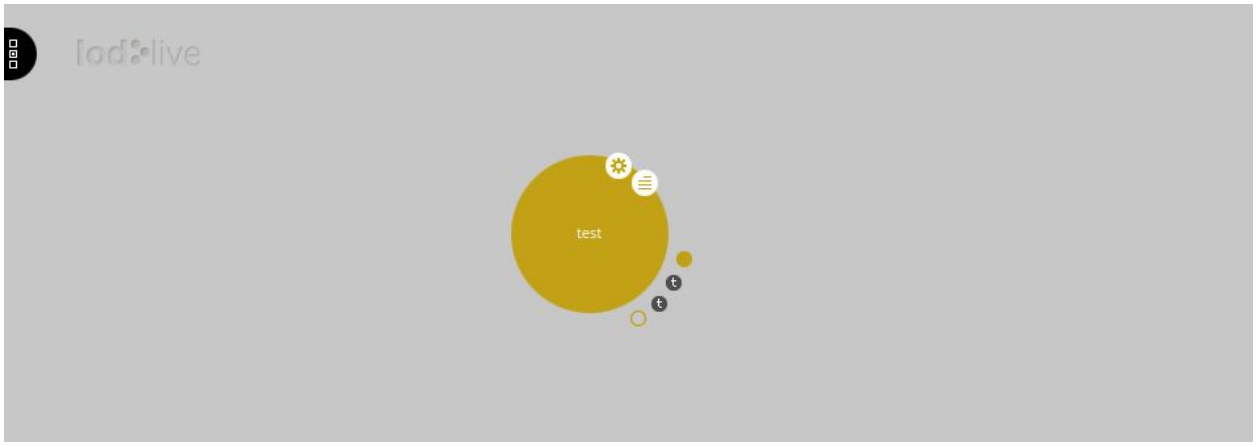


Fig. 11: Graph representation of your Latin text in LiLa. You can now explore your text in relation to all other information already contained in the LiLa Knowledge Base by clicking on all of the “satellite” nodes.

Let us now query our Latin text! Open the LiLa query interface at https://lila-erc.eu/sparql/ (Fig. 12).

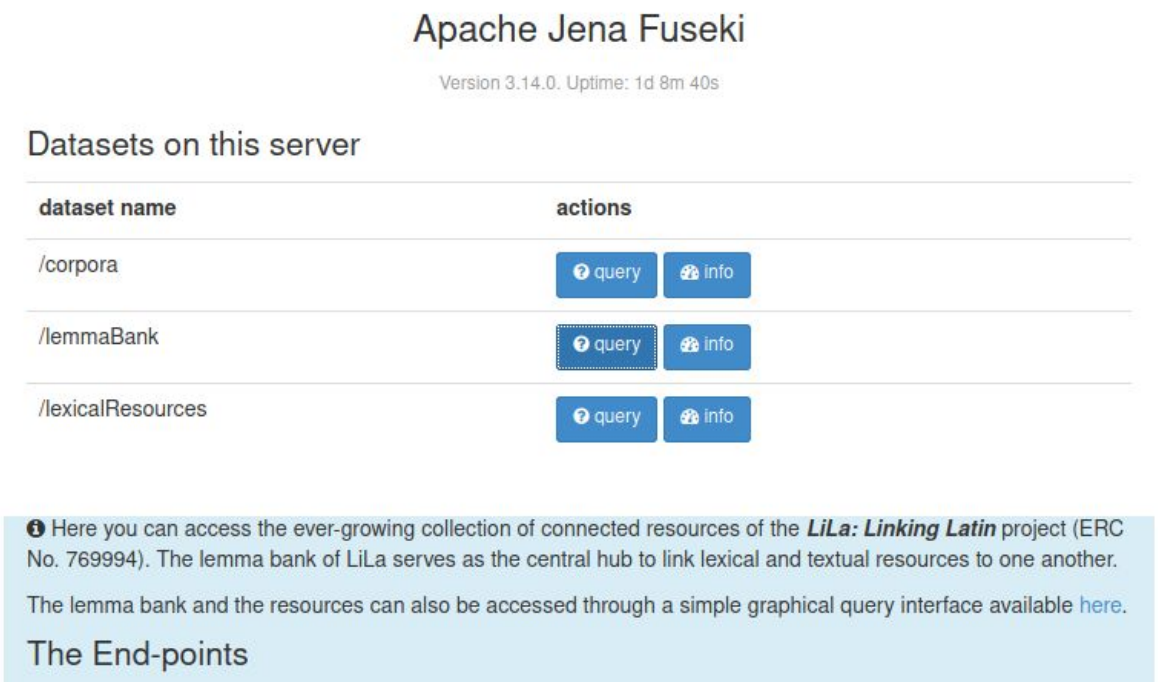


Fig. 12: LiLa’s main SPARQL interface.

Look for the `lemmaBank` dataset and click on the QUERY button next to it. You should now see this SPARQL query interface (Fig. 13):

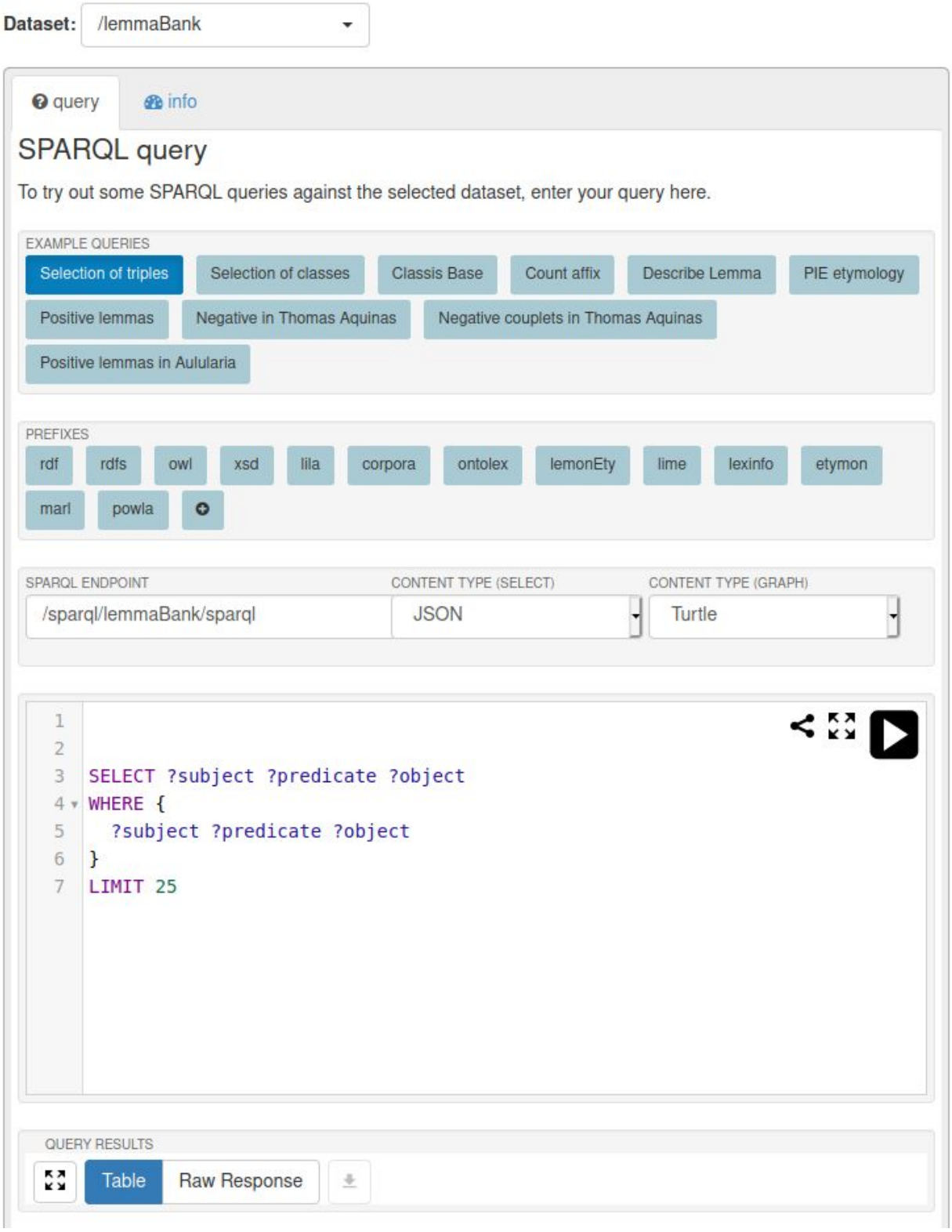


Fig. 13: SPARQL interface to query the LiLa lemma bank.

From the query file `LiLa_SPARQL_queries.rq` provided as part of this activity's materials, copy the first SPARQL query and paste it in this SPARQL interface box (Fig. 14). As explained in that file, this query can be used to produce the list of all lemmas in our text at least once in decreasing order of frequency.

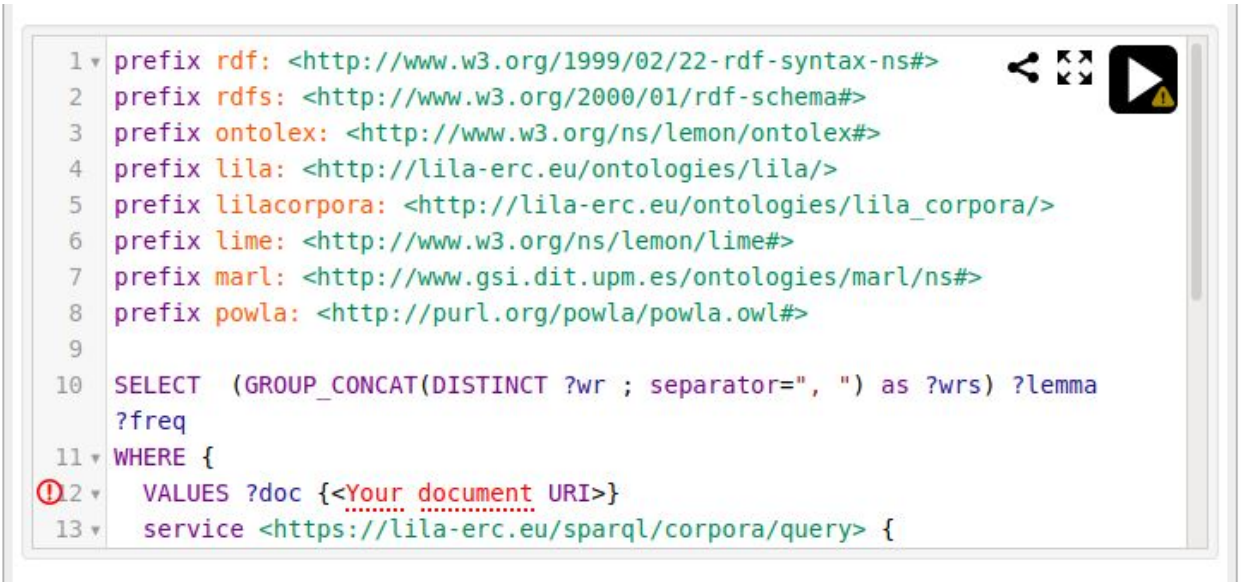
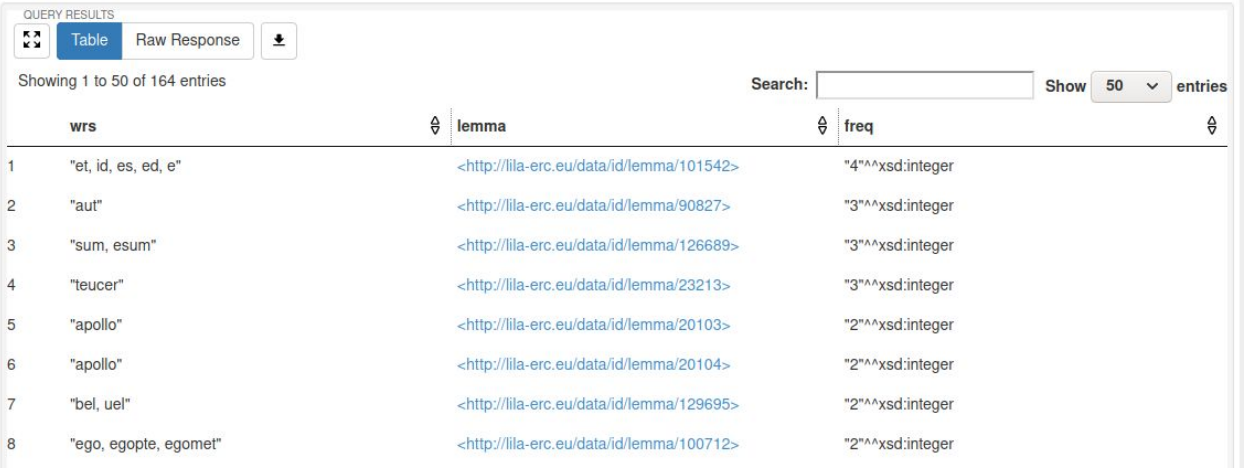


Fig. 14: Pasting a SPARQL query in the LiLa SPARQL interface.

Doing so will flag-up a warning, visible thanks to the (!) icon. What the system is asking you is to insert the URI of your document (that displayed in Fig. 9). Paste that URI in between the angle brackets (< >) provided. Now click on the large black/white play button to the right of your screen to run the query on the Knowledge Base! The results of your query will be shown below the query box (Fig. 15).



	wrs	lemma	freq
1	"et, id, es, ed, e"	<http://lila-erc.eu/data/id/lemma/101542>	"4"^^xsd:integer
2	"aut"	<http://lila-erc.eu/data/id/lemma/90827>	"3"^^xsd:integer
3	"sum, esum"	<http://lila-erc.eu/data/id/lemma/126689>	"3"^^xsd:integer
4	"teucer"	<http://lila-erc.eu/data/id/lemma/23213>	"3"^^xsd:integer
5	"apollo"	<http://lila-erc.eu/data/id/lemma/20103>	"2"^^xsd:integer
6	"apollo"	<http://lila-erc.eu/data/id/lemma/20104>	"2"^^xsd:integer
7	"bel, uel"	<http://lila-erc.eu/data/id/lemma/129695>	"2"^^xsd:integer
8	"ego, egopte, egomet"	<http://lila-erc.eu/data/id/lemma/100712>	"2"^^xsd:integer

Fig. 15: Results of our SPARQL query.

You can now repeat this process with all other queries provided in the `LiLa_SPARQL_queries.rq` file!

Questions

- Q: At this moment, diachronic linking of variants is considered and, if so, how it made?
    - A: Not at the moment, no. Definitely possible, but would require a significant amount of work :)
- Q: What would be the abstract instance showing to both lemmas spectrum/spectrom instances?
    - A: In LiLa, its URI
    - A: *Spectrum* is not currently among the ca. 9,300 lemmas of the Latin WordNet but we can certainly attach synsets/concepts to it in the next LiLa release, e.g. http://wordnet-rdf.princeton.edu/lemma/spectre
- Q: For long texts, can the work of choices in TextLinker be saved and later reloaded so that someone can close the browser and continue later? (Or could be added as next feature?)
    - A: TO BE IMPLEMENTED.
- Q: If we provide OCR versions of Mediaeval Latin editions of - say - biblical books, would the lemmatiser take into account variant spellings, and so on?
    - A: That depends on the data we used to train the pos-tagger/lemmatiser. The more data we train the lemmatiser on, the more likely it is to recognise variants.
    - A: If you mean OCR *errors*, if these are not corrected the lemmatiser won't be able to parse the text.
- Q: As the Latin texts in your interoperable network have been translated into vernacular languages, do you see any future linking possible with vernacular sources? Would you see yourselves partnering up with someone in such an endeavour?
    - A: That would be wonderful, yes! The Ancient Greek loanword resource that we described in our presentation points in that direction. But in order to link LiLa to other languages it would be necessary for other languages to produce a linguistic linked data Knowledge Base. If, for instance, we had a LiLa equivalent for Ancient Greek, we could connect "our" Ancient Greek words (e.g. *apsinthion*, as shown in our presentation) to those collected in the Ancient Greek Linked Data Knowledge Base.
- Q: Maybe I missed it: what happens if the token is not given the proper UPOS when the text is processed? For instance I had the word "Occidam" classified as "Lemma: Occida - UPOS: PROPN" but it is not a proper noun, it is a verb!
    - A: In that case, click on OCCIDAM and then, on the right of the screen, on the "unlink" ICON. Doing so will give you the option of searching for and choosing the correct lemma in the LiLa Knowledge Base.
    - I linked it to the verb (occido — LodView (lila-erc.eu)), but it is still shown as a noun..

- 

    - I think that might be a bug on our end, then. The system should have correctly linked your lemma to the VERB but the text does reflect this change. We'll work on it, thanks for flagging this up! :) OK, thanks!

Issues/Requests

- **Problem 1: On some computers, the LodLive window returns a 404 error.**
    - Diagnosis: this seems to be a problem with browsers set to French.
    - Workaround: participants using browsers set to French should change their browser language to English.
- **Problem 2: Undo action in TextLinker**
- **Problem 3: Add the possibility to check the lemmas in the LiLa KB when the drop-down menu appears**
- **Problem 4:** Once an incorrectly lemmatised token has been corrected, the TextLinker doesn't display the updated lemma and PoS but the original (incorrect) information.
- Nice to have : in the lod graph tool, for example exploring albus, and having opened its base node "Base205", when going over the satellite dots of the base it could be useful to read in the tooltip also the actual lemma (currently it only shows the lemma id and the type of relations); otherwise one has to check/open all the satellite dots to discover where are the subnodes of the base he's searching
- Nice to have : in the lod graph tool it could also be useful to have an option to close with one click all the currently open derived nodes of a base. it seems possible for the satellite dots (reclicking the cogwheel dot), but if someone has opened 20 descendants of a base it looks like he must close each of them manually
- Nice to have : also a kind of textbox filter near the base cogwheel, in adjunction to the satellite dots, to search quickly the descendants of a base (e.g. those starting with "sub…") and to open them immediately
- As it was mentioned during the workshop, the disambiguation is the most time-consuming part, and sometimes it is quite difficult to understand how two lemmas differ (we have to follow several links, etc.). I was wondering if it was a possible solution to add a translation, e.g. by linking to the Perseus dictionary? Or is this too complicated?
    - As mentioned in our presentation, we're in the process of cleaning and adding the Latin WordNet to LiLa. This means that the TextLinker will also provide word meanings to help with disambiguation :) Word meanings will be added as the cleaning of the Latin WordNet progresses.

Latin text suggestions

Text	URL	Comment
Pliny the Elder, Naturalis Historia	https://penelope.uchicago.edu/Thayer/L/Roman/Texts/Pliny_the_Elder/2*.html	
I'm interested in David Hilchen text corpora but unfortunately I cannot provide any URL's. Given that, I would prefer any neo-latin text	http://www.mlat.uzh.ch/MLS/xanfang.php?corpus=9&lang=0	
Petrus Ransanus: Epitome Rerum Hungararum	https://szovegtar.iti.mta.hu/hu/muvek/epithoma-rerum-hungararum-id-est-annalium-omnium-temporum-liber-primus-et-sexagesimus/	Copy-pasting from the PDF works well
De Bello Gallico	http://www.latin.it/autore/cesare/de_bello_gallico/!01!liber_i	
Thomas de Vio Cardinal Cejetan texts (1462-1534), on analogy, biblical exegesis, commentaries to Aquinas S. Th. Etc.		Can't find a link :(
Ursus Beneventanus	https://github.com/paolomonella/ursus	
The Book Psalms/mediaeval source material for vernacular translations	http://www.mlat.uzh.ch/MLS/xanfang.php?tabelle=Biblia_cps0&corpus=0&id=Biblia_cps0,%20Biblia%20Sacra%20iuxta%20Vulgatam%20Clementinam	
Giovanni Pico della Mirandola, Conclusiones secundum Thomam	http://cds.library.brown.edu/projects/pico/about.php	
Hor., carm. VII	http://mizar.unive.it/mqdq/public/testo/testo?codice=HOR%7Ccar2%7C007	Book? :)
The Vulgate (Psalms, Canticles, and New Testament for starters?)	http://music2.princeton.edu/chant_html/bibles.html	This gathers several online texts (provided via links), but i'm not sure about the possibility of distribution and/or "text aspiration"

Feedback

A space to let us know how we can improve, be that the pace of the activity or a feature request for the Text Linker! Anything 😊

Some random comments from Joel:

- You have my sincere thanks and congratulations. The amount of work you've put into your datasets is impressive and exhausting to think about! I also appreciate the contagious enthusiasm exhibited by all members of the team.
- There is an initiative underway, by Christian Chiarcos, to do better in LOD linguistic annotations: https://github.com/ld4lt/linguistic-annotation . The work is in its early days, and I recommend participation.
- In the coming year or two I hope to refine the TAN-LM and TAN-A-mor formats, to try to show that it can be treated as a kind of highly regulated RDF serialization. It might be a complete failure! Whether it is or not, I look forward to trying to synchronize my efforts with you all.

Some random questions from Joel (you can contact me at kalvesmaki@gmail.com or director@textalign.net):

- For our test example, there was *carmine*, with ambiguous lemmata. But they looked identical to me, despite having two different lemmata numbers. Was my eyes playing tricks on me?
- In our wonderful quiz we learned that there are three lemmata for *cum*. But in the sample Horace text, no cases of *cum* resulted in ambiguities. Why? I'm curious to know if you've developed a weighted scoring system, and how that worked out.
- In the Text Alignment Network, I have developed functions for Morpheus's morphological service for Latin and Greek, to help accommodate those who are working within XSLT to incorporate queries straight into their XML workflow, to develop RDF-friendly lexico-morphological data for their TEI files. These are successful because I can pass a URL to the API and get results. Can I can a URL to LiLa, something like this: https://lila-erc.eu/sparql/corpora/query?SELECT%20%2A%0AWHERE%20%7B%0A%3Fs%20%3Fp%20%3Fo%20.%0A%7D%0ALIMIT%203
  If I can, then I can write a function in the TAN library that taps into LiLa.
- Both my TAN work has made me realize that straightforward triples are fine for simple straightforward linguistic annotation, but not for more complex cases. For this, it seems that there needs to be reification, and adoption of RDF*. What complications if any would there be if I adapted and exposed LiLa RDF in this more reified version? Have you given any thought to moving your triples toward this model?