# Towards a Representation of *Citations* in Linked Data Lexical Resources

Fahad Khan, Federico Boschetti
ILC-CNR, Pisa
{firstname.secondname}@ilc.cnr.it

# Introduction

Given the current popularity of **linked data** as a means of publishing datasets and the move towards making more **lexical resources** available on the **Semantic Web** (promoted esp. by the **elexis project**), it is worth asking:

*How can we best exploit the technical possibilities that linked data offers in order to model the different kinds of phenomena that we find in lexicographic datasets -- including **retrodigitised dictionaries**?*

A related question is: *what can linked data offer us that other approaches, such as **TEI**, cannot?*

# Introduction

Although the questions in the last slide are much too broad for me to be able to deal with properly in the space of this talk **in general** I do want to zoom in on a particular kind of lexicographic phenomenon, **the use of citations and attestations**, and look at how to we can model it in linked data and the sorts of advantages that might bring.

I intend to discuss the ways in which **linked data**, and especially semantic web languages like **RDFS** and **OWL**,  allow us to pursue a more traditional knowledge engineering approach w/r/t to citations and attestations *a la* classical AI. I want to argue that linked data offers a more natural way of modelling **networks of relations** between **heterogeneous** datasets.

# Introduction

In this talk I will look at two **case studies** which will allow me to raise a number of interesting questions w/r/t the correct modelling of citations/attestations. The first example is from the **Liddell-Scott-Jones Ancient Greek-English** lexicon, the second is the treatment of an **Italian homonym *riprovare*** in two different Italian dictionaries.

At the end of my talk I will propose a minimal vocabulary for attestations based on the **Ontolex-Lemon** model which I have named ***lemonBib***.

**First however** -- and in order that we're all on the same page -- I want to give a quick resume of some important **linked data fundamentals** which will be relevant to what follows.

# Some (Basic) Definitions

**Linked Data**: a method of publishing **structured data** so that it can be **interlinked** and become more useful through **semantic queries**.  (*Source: Wikipedia*)

**The Semantic Web**: a **web of datasets** that are structured and linked together using **a common set of standards and technologies** so that they can be **more easily processed by computers**.

*Linked Data is one very important way of making the semantic web a reality.*

# Some More Definitions

The **Resource Description Framework** (RDF) is the standard way of modeling data on the Semantic Web. Importantly RDF makes the constraint that *we describe our data using <u>only</u> statements (called triples) of the form*:

**Subject- Predicate-Object**

Where the Subject, Predicate and Object are each resources  ('conceptual things') with their own **Uniform Resource Identifiers** (URIs) and we can access data on them using the standard **HTTP protocol;** the protocol used to access web pages on the internet. Unlike HTML webpages, though, RDF datasets are intended for machine rather than human consumption.

# RDF Graphs v. TEI trees

RDF has the **big advantage** that it is naturally suited to describing data which is structured in the form of **graphs/networks**. It also makes it simple to link to other arbitrary (**heterogeneous**) datasets using formal languages (like **RDFS** and **OWL**) to describe the meanings of the links between datasets -- as well as the classes used as categories in the data itself.

On the other hand **TEI** is much more suited to describing **tree structures** and gives us much less scope for making a formal description of the semantics of elements/creating typed links between datasets. Up till now this was regarded as sufficient for creating digital editions. **But e.g., citational/etymological information is less easily encoded as a tree.**

# In Summary

It is clear that given that they involve **bibliographic, temporal, and other kinds of hetereogeneous sources** lexicographic **citations** offer a promising case study for showing the benefits of graph-based linked data modelling in publishing lexica.

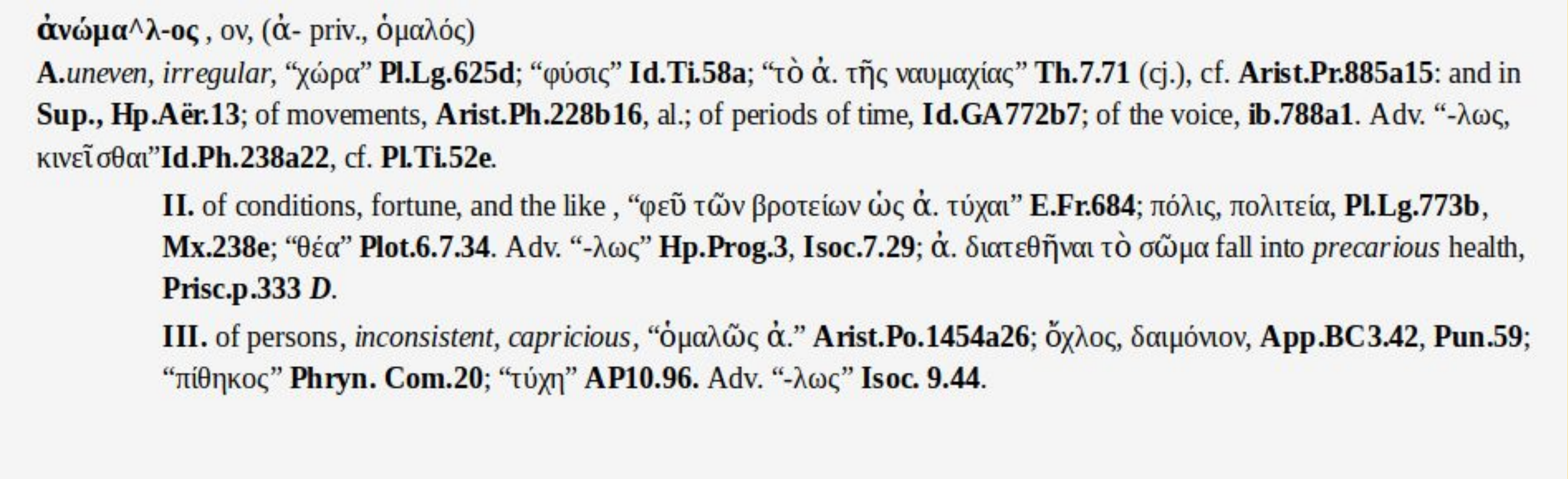
I want to argue however that modelling lexical citations properly **requires us to make a distinction between attestations and citations**. This is the kind of distinction that the use of RDF seems to make more visible.

# An Anomalous Example

Citations can be used to *attest* various different properties of a lexical entry, e.g., **orthographic**, **semantic**, **phonetic**. But they can also be used for other purposes. Our first example shows this and also sheds some light on how attestations and citations are related together.

We will look at the entry for **ἀνώμᾰλος** (anomalos) from the hugely influential Liddell-Scott-Jones ancient Greek-English lexicon (made available online by the **Perseus project**).

# An Anomalous Example



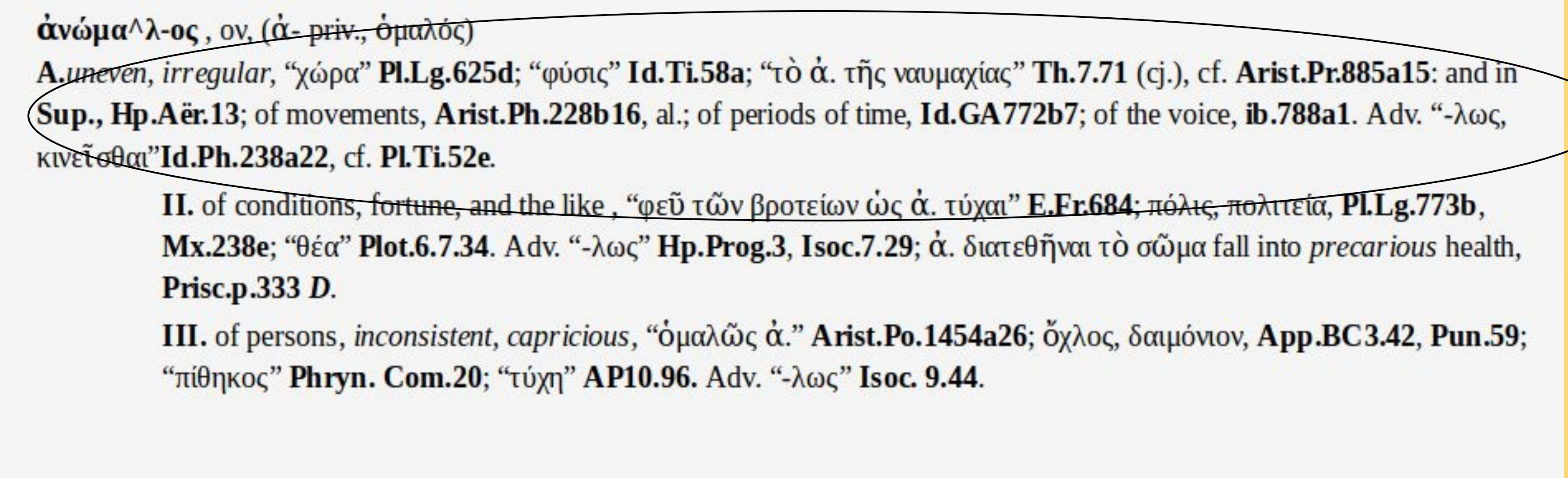
**ἀνώμα^λ-ος** , ον, (ἀ- priv., ὁμαλός)

**A.** *uneven, irregular*, “χώρα” **Pl.Lg.625d**; “φύσις” **Id.Ti.58a**; “τὸ ἀ. τῆς ναυμαχίας” **Th.7.71** (cj.), cf. **Arist.Pr.885a15**: and in **Sup., Hp.Aër.13**; of movements, **Arist.Ph.228b16**, al.; of periods of time, **Id.GA772b7**; of the voice, **ib.788a1**. Adv. “-λως, κινεῖσθαι”**Id.Ph.238a22**, cf. **Pl.Ti.52e**.

**II.** of conditions, fortune, and the like , “φεῦ τῶν βροτείων ὡς ἀ. τύχαι” **E.Fr.684**; πόλις, πολιτεία, **Pl.Lg.773b**, **Mx.238e**; “θέα” **Plot.6.7.34**. Adv. “-λως” **Hp.Prog.3**, **Isoc.7.29**; ἀ. διατεθῆναι τὸ σῶμα fall into *precarious* health, **Prisc.p.333** *D*.

**III.** of persons, *inconsistent*, *capricious*, “ὁμαλῶς ἀ.” **Arist.Po.1454a26**; ὄχλος, δαιμόνιον, **App.BC3.42**, **Pun.59**; “πίθηκος” **Phryn. Com.20**; “τύχη” **AP10.96.** Adv. “-λως” **Isoc. 9.44**.

# An Anomalous Example



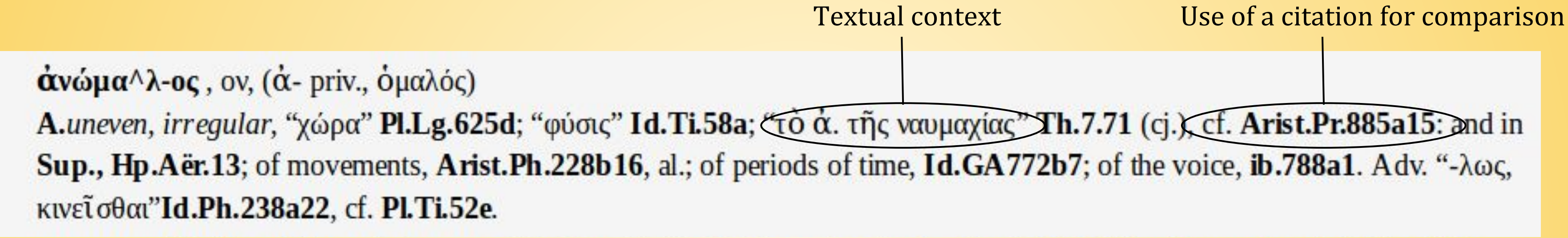
**ἀνώμα^λ-ος** , ον, (ἀ- priv., ὁμαλός)

**A.** *uneven*, *irregular*, “χώρα” **Pl.Lg.625d**; “φύσις” **Id.Ti.58a**; “τὸ ἀ. τῆς ναυμαχίας” **Th.7.71** (cj.), cf. **Arist.Pr.885a15**: and in **Sup., Hp.Aër.13**; of movements, **Arist.Ph.228b16**, al.; of periods of time, **Id.GA772b7**; of the voice, **ib.788a1**. Adv. “-λως, κινεῖσθαι”**Id.Ph.238a22**, cf. **Pl.Ti.52e**.

**II.** of conditions, fortune, and the like , “φεῦ τῶν βροτείων ὡς ἀ. τύχαι” **E.Fr.684**; πόλις, πολιτεία, **Pl.Lg.773b**, **Mx.238e**; “θέα” **Plot.6.7.34**. Adv. “-λως” **Hp.Prog.3**, **Isoc.7.29**; ἀ. διατεθῆναι τὸ σῶμα fall into *precarious* health, **Prisc.p.333** *D*.

**III.** of persons, *inconsistent*, *capricious*, “ὁμαλῶς ἀ.” **Arist.Po.1454a26**; ὄχλος, δαιμόνιον, **App.BC3.42**, **Pun.59**; “πίθηκος” **Phryn. Com.20**; “τύχη” **AP10.96**. Adv. “-λως” **Isoc. 9.44**.

# An Anomalous Example



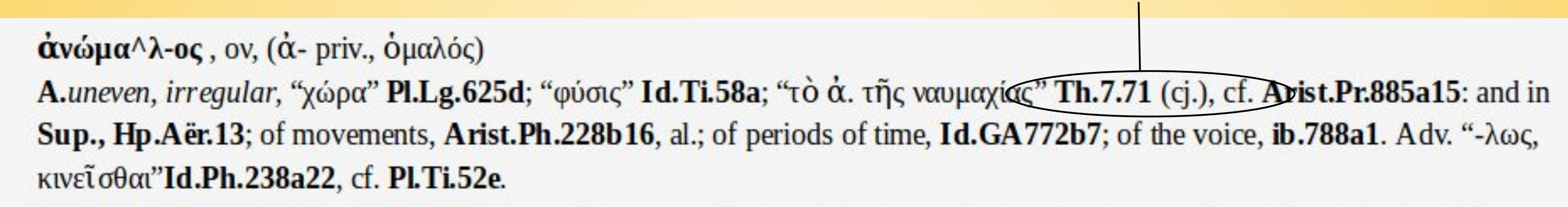
Textual context

Use of a citation for comparison

**ἀνώμα^λ-ος** , ον, (ἀ- priv., ὁμαλός)
**A.***uneven*, *irregular*, “χώρα” **Pl.Lg.625d**; “φύσις” **Id.Ti.58a**; “τὸ ἀ. τῆς ναυμαχίας” **Th.7.71** (cj.), cf. **Arist.Pr.885a15**: and in **Sup., Hp.Aër.13**; of movements, **Arist.Ph.228b16**, al.; of periods of time, **Id.GA772b7**; of the voice, **ib.788a1**. Adv. “-λως, κινεῖσθαι”**Id.Ph.238a22**, cf. **Pl.Ti.52e**.

Most of the citations in the example are used to *attest* to different shades of meaning of the word in question, with the **textual context** of an attestation **explicitly given** in one case.  In other cases citations are used to contrast with other citations: without necessarily attesting to the word sense being dealt with. This use of the citation is annotated by the abbreviation ‘**cf.**’.

# An Anomalous Example



Conjectural citation

**ἀνώμα^λ-ος** , ον, (ἀ- priv., ὁμαλός)

**A.** *uneven, irregular*, “χώρα” **Pl.Lg.625d**; “φύσις” **Id.Ti.58a**; “τὸ ἀ. τῆς ναυμαχίας” **Th.7.71** (cj.), cf. **Arist.Pr.885a15**: and in **Sup.**, **Hp.Aër.13**; of movements, **Arist.Ph.228b16**, al.; of periods of time, **Id.GA772b7**; of the voice, **ib.788a1**. Adv. “-λως, κινεῖσθαι”**Id.Ph.238a22**, cf. **Pl.Ti.52e**.

It is also interesting to note that one of the citations, **‘Th.7.71’**, is marked with a **‘(cj.)**’ meaning that it is conjectural -- i.e., it is based on **a reconstruction of the original text**. In this case we can say that the entry cites the text (from the corpus of works attributed to Thucydides) even though the original text might not have actually attested the sense itself.

# If at first you don't succeed...

Our second example involves **Dante Alighieri**. It is intended to show how two authoritative lexical sources can **disagree** on the *meaning* of a citation. It revolves around the following two Italian homonyms:

- *riprovare* 'to try something again' (from *provare* 'to try' and the prefix *ri-* which adds the sense of repetition); call this **riprovare1.**
- *riprovare* 'to scold, rebuke' (in this sense it is cognate with the English verb *reprove*); call this **riprovare2**.

# If at first you don’t succeed

- The **motto** of the 16th century ***Accademia del Cimento*** “provare e riprovare”, *try and try again*, captured the spirit of scientific endeavour promoted by that organisation. I.e., **riprovare1** *is attested by* **the AdC motto**
- **Dante’s Paradiso** (Par. III, 1-3) contains a passage *attesting* to **riprovare2**, i.e.,
  - ‘Quel sol che pria d’amor mi scaldò 'l petto,  
    di bella verità m'avea scoverto,  
    provando e riprovando, il dolce aspetto’  
    (*That Sun, which erst with love my bosom warmed/ Of beauteous truth had unto me discovered/By proving and reproving, the sweet aspect*.)

# Treccani v. Battaglia

The two authoritative Italian-language lexicographic resources, *il vocabolario Treccani* and *il Grande Dizionario della Lingua Italiana* (GDLL) treat these homonyms and the previous sources as follows:

- *Treccani's entry for **riprovare1** cites the AdC motto as attesting to the entry (see [http://www.treccani.it/vocabolario/riprovare1](http://www.treccani.it/vocabolario/riprovare1))*
- *Treccani's entry for **riprovare2** cites Par. III, 1-3. as attesting to the entry (see [http://www.treccani.it/vocabolario/riprovare2](http://www.treccani.it/vocabolario/riprovare2))*
- *GDLL's entry for **riprovare1** cites Par. III, 1-3. and the AdC motto as attesting to the entry*

# Attestations and Citations

There are a number of truth claims here that we can list as follows:

1. Treccani’s entry for **riprovare1** cites the AdC motto
2. Treccani’s entry for **riprovare2** cites Par. III, 1-3.
3. GDLL’s entry for **riprovare1** cites Par. III, 1-3.
4. **riprovare1** is attested by Par. III, 1-3.
5. **riprovare2** is attested by Par. III, 1-3.
6. **riprovare1** is attested by AdC

# Attestations and Citations

There are a number of truth claims here that we can list as follows:

1. Treccani's entry for **riprovare1** cites the AdC motto
2. Treccani's entry for **riprovare2** cites Par. III, 1-3.
3. GDLL's entry for **riprovare1** cites Par. III, 1-3.
4. **riprovare1** is attested by Par. III, 1-3.
5. **riprovare2** is attested by Par. III, 1-3.
6. **riprovare1** is attested by AdC

Statements 1-3 describe *citations* at the level of bibliography.

# Attestations and Citations

There are a number of truth claims here that we can list as follows:

1. Treccani's entry for **riprovare1** cites the AdC motto
2. Treccani's entry for **riprovare2** cites Par. III, 1-3.
3. GDLL's entry for **riprovare1** cites Par. III, 1-3.
4. **riprovare1** is attested by Par. III, 1-3.
5. **riprovare2** is attested by Par. III, 1-3.
6. **riprovare1** is attested by AdC

Statements 4-6 describe *attestations* at what we might call a *lexical* level.

# Attestations and Citations

There are a number of truth claims here that we can list as follows:

1. Treccani's entry for **riprovare1** cites the AdC motto
2. Treccani's entry for **riprovare2** cites Par. III, 1-3.
3. GDLL's entry for **riprovare1** cites Par. III, 1-3.
4. **riprovare1** is attested by Par. III, 1-3.
5. **riprovare2** is attested by Par. III, 1-3.
6. **riprovare1** is attested by AdC

Statement 3 is true, but its corresponding lexical claim, its related truth content,  Statement 4 is false. Both these levels may be interesting independently of one another.

# What have we learned?

- Citations are **not** always used to **attest to the lexical entry/sense in which they're included;**
- Citations can target **conjectural reconstructions** and thus may have an uncertain status that is annotated in the text of the entry itself with cj;
- What I call attestations can be viewed as the conceptual content of citations when they are used as direct evidence of a certain lexical property having been used in a corpus;
- Citations can be interesting even when they don't correspond to attestations or when their corresponding attestations are incorrect or dubious.

# Modelling Citations and Annotations

By forcing us to **explicitly model our data** in terms of Subject-Predicate-Object triples RDF encourages us to think in terms of simple **declarative truth claims**: i.e., they make the preceeding considerations more salient. This is even more true wrt RDFS and OWL as these are much more expressive formal languages (OWL is a of description logic) and enable us/encourage us to make the meanings of our data much more ‘**explicit**’

The advantage of making this distinction is that it makes these different kinds of information more easily findable and queryable using the Semantic Web Query Language **SPARQL** for example.

# TEI's Views on Lexical Data

It will be useful to look at the distinction that the TEI Dictionary guidelines make between the three different views on lexical data:

a) The **typographic view** concerns the layout of a page, e.g., where the line breaks are in the text and how the entries are arranged on any single page.

b) The **editorial view**:  which words are used and in which order along with the exact placement of punctuation in each entry.

c) The **lexical view**, relates to the conceptual, linguistic, content of a lexicon and each of its individual entries

# TEI’s Views on Lexical Data

The TEI-Dict guidelines admit that the TEI-Dict model itself deals with all three levels. However it is difficult/extremely verbose to deal with the first and to a large extent the second level in RDF. This suggests a division of labour between the two models. The **ontolex-lemon** model is (mostly) concerned with the third TEI level. This is the same level at which attestations should be modelled, e.g.,

- ***riprovare1*** *is a verb*
- ***riprovare1*** *can be translated by ‘to try again’*
- ***riprovare1*** *is attested by the motto of the Accademia del Cimento*

It is at this level that our proposed vocabulary for attestations operates.

# Different Views

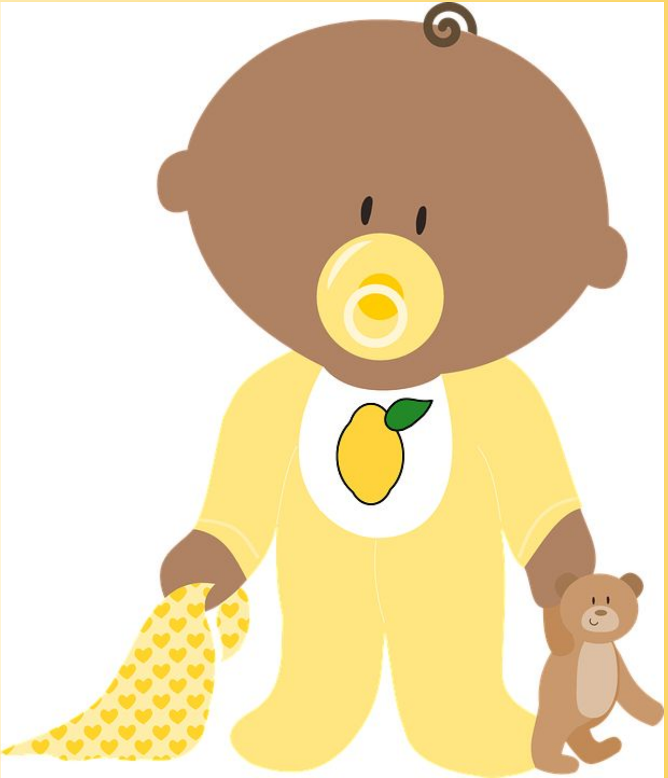
Citations are at a different level from attestations. Rather they are **speech acts** that can be used for different purposes but seem to be mainly used for stating attestation relations between texts and lexical properties.  They are part of the bibliographic/metadata level of description of a text, e.g.,

- *Treccani is a monolingual Italian dictionary*
- *Treccani is published by the Istituto dell'Enciclopedia Italian publishing house*
- *Treccani is available online*
- *Treccani's entry for* ***riprovare1*** *cites the motto of the Accademia del Cimento*

This level is already dealt with by other dedicated vocabularies such as CITO

# lemonBib

- A fairly minimal vocabulary that places the class **`Attestation`** at its centre
- meets the following requisites:
  - Relates attestations to their corresponding citations and vice versa;
  - Relates an attestation to the text to which it refers;
  - Allows the specification of the textual context of an attestation + its precise location;
  - Allows the conjectural status of an attestation to be specified
- *http://lari-datasets.ilc.cnr.it/lemonBib*



# lemonBib - features

- **`Attestation`**: a class that consists of individuals that reifies the relationship of lexical attestation between an element in a lexicon and a bibliographic element
- **`attests`**: an object property that relates an individual of the class Attestation with a lexical element
- **`isAttestedBy`**: an object property that relates a lexical element with an individual of the class Attestation
- **`involvedinAttestation`**: an object property that relates a citation with an Attestation
- **`hasContext`**: a datatype property that relates an Attestation with a string
- **`foundIn`**: an object property that relates an Attestation with a location in a text where it can be found

*Caveat: There is currently discussion on an extension for attestations in the ontolex-lemon model.*

*Parts of lemonBib may or may not be adopted into this ontolex extension*

# Conclusions

- Lexicography is *arguably* at the interface between **digital humanities** and **computational linguistics**. We can be interested in dictionaries at different levels and for different reasons, historical, bibliographical, cultural, *as well as* in terms of their purely linguistic content;
- It is probably not  a good idea prioritise one of these levels to the exclusion of the others (c.f. the **Nenufar project** which tracks the treatment of words over different 20th cenury editions of ***Le Petit Larousse***);
- Linked Data/Semantic Web standards enable us to make the information at these different levels more explicit and more accessible.