COST STSM Report
COST-STSM-IS1305-35141
Host: Eveline Wandl-Vogt
Enhancement of a visual analysis tool for historical dictionaries

Alejandro Benito Santos
Universidad de Salamanca
Start date: 14/11/16
End date: 18/11/16

1 Purpose of the STSM

Before the STSM, a team of experts in visualization at University of Salamanca, in collaboration with the Digital Humanities Center at the Austrian Academy of Sciences, created a visual analysis tool to interactively explore data coming from the Wrterbuch der bairischen Mundarten in sterreich (WB) historical dictionary. During the process of conception of such tool, several other computational methods were applied apart from purely visualization techniques. The prototype tool was recently presented at the TEEM Conference 2016 celebrated in Salamanca (Spain)[4]. The underlying data format that the tool supported was TUSTEP-XML. Ad-hoc importers were developed that adapted and imported such textual data into the documental search engine at the core of the tool. Among the future lines of research that were pointed out at the end of the paper were those related to the support of other data formats, with the ultimate aim of connecting different lexicographic sources. Recently, experts at the Center for Digital Humanities at The Austrian Academy of Sciencies converted the TUSTEP-XML data to the more general TEI standard. This will allow import/export of the data to external sites like `http://dictionaryportal.eu` or others. Another RDF data set from the Biodiversity and Linguistic Diversity project (also connected to COST-ENeL)[5] was also analyzed. RDF is the standard used in the Semantic Web, and it is key to make cultural/conceptual connections from lexicographic data and will be the format used by the prototype in order to connect concepts between different data sets.

During the STSM, one of the main aims was for the data visualizer to get acquainted with the new format and its particularities, as well as discovering new processes and key artifacts used by lexicographers during their work, so the tool can better adapt to the actual needs of the research processes. All these needs were at the end extracted in the shape of user stories that adapt to the AGILE software development methodology that is being used in the prototype tool. The process of requirement extraction is a long and involves a lot of discussion between the different stakeholders of the project. In this case it was accelerated by means of supportive software and live coding sessions that will be discussed later in this document.

2 Investigation/Research

A series of steps were followed throughout the scientific mission:

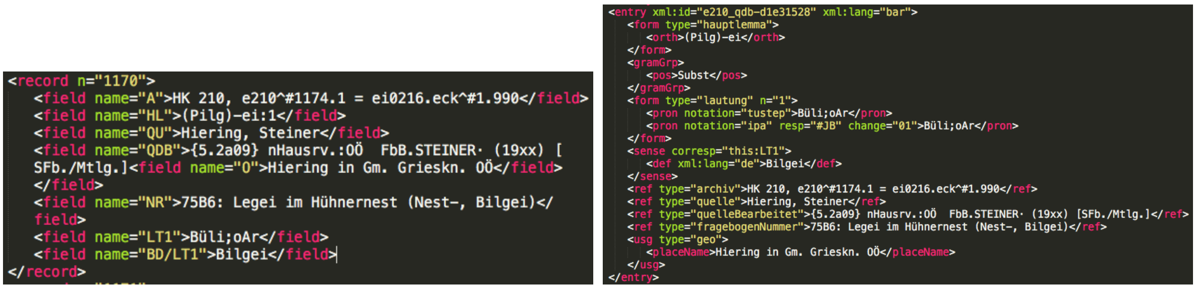
1. Understanding particularities of the RDF and TEI formats
2. Analyzing artifacts and work flows used by lexicographers in their daily work
3. Live analysis/coding sessions
4. Extracting user stories from 2 and 3.

The first two points are related to knowledge acquisition by the data visualizer needed to orient the following sessions related to points 3 and 4. The first one is linked to the data format that is used by the academy. The sessions conducted in point 2 are key to understand the work flows of the lexicographers and thus to create meaningful visualization systems that can help them in their daily work. In the following sections we comment each of the sessions related to each point, and the results obtained at the end.

3 Data Formats

3.1 TEI

As it was previously stated in this document, historical dictionary information has been migrated to the TEI standard. This format differs substantially from the previously used TUSTEP-XML format, and we comment these differences in this section. A first overview is given the following image, which illustrates these differences:



```
<record n="1170">
    <field name="A">HK 210, e210^#1174.1 = ei0216.eck^#1.990</field>
    <field name="HL">(Pilg)-ei:1</field>
    <field name="QU">Hiering, Steiner</field>
    <field name="QDB">{5.2a09} nHausrv.:OÖ  FbB.STEINER· (19xx) [
    SFb./Mtlg.]<field name="O">Hiering in Gm. Grieskn. OÖ</field>
    </field>
    <field name="NR">75B6: Legei im Hühnernest (Nest-, Bilgei)</
    field>
    <field name="LT1">Büli;oAr</field>
    <field name="BD/LT1">Bilgei</field>
</record>
```

```
<entry xml:id="e210_qdb-d1e31528" xml:lang="bar">
    <form type="hauptlemma">
        <orth>(Pilg)-ei</orth>
    </form>
    <gramGrp>
        <pos>Subst</pos>
    </gramGrp>
    <form type="lautung" n="1">
        <pron notation="tustep">Büli;oAr</pron>
        <pron notation="ipa" resp="#JB" change="01">Büli;oAr</pron>
    </form>
    <sense corresp="this:LT1">
        <def xml:lang="de">Bilgei</def>
    </sense>
    <ref type="archiv">HK 210, e210^#1174.1 = ei0216.eck^#1.990</ref>
    <ref type="quelle">Hiering, Steiner</ref>
    <ref type="quelleBearbeitet">{5.2a09} nHausrv.:OÖ  FbB.STEINER· (19xx) [SFb./Mtlg.]</ref>
    <ref type="fragebogenNummer">75B6: Legei im Hühnernest (Nest-, Bilgei)</ref>
    <usg type="geo">
        <placeName>Hiering in Gm. Grieskn. OÖ</placeName>
    </usg>
</entry>
```

Figure 1: Comparison of the same lemma entry in the TUSTEP (left) and TEI (right) format representations

New TEI importers for ElasticSearch[2] were developed on-site during the first days of the STSM. This step was needed for the knowledge-exchange sessions that came the following days.

3.2 RDF

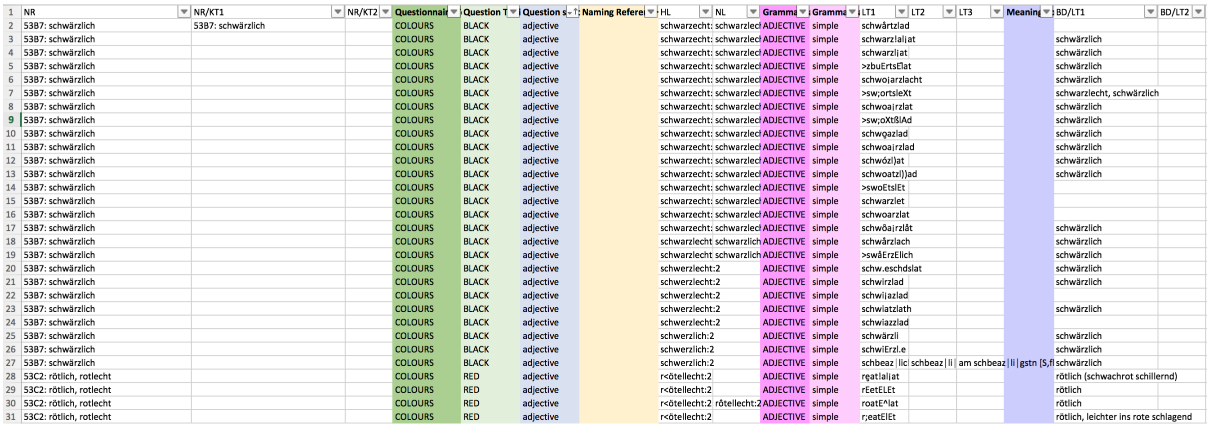
An example subproject was also analyzed during the STSM. This service links scientific names of certain plants to their common names in different languages. This is a good example on how to generate cultural connections between different languages and dialects, and will serve as the basis of graph visualizations in the prototype.

4 Artifact Analysis

Before moving to the coding sessions, it was necessary to understand the nature of the work lexicographers at the Academy are performing. In Digital Humanities there is usually a big gap between the different participants in a projects, who are usually related to different areas of expertise (In our case data visualization and lexicography). Effectively reducing these gaps involves a great deal of time and effort and it is one of the biggest challenges of the project.

One of the main aims of the research is to visualize cultural connections and expose the results to third parties. As of now, the process of textual feature extraction is completely handmade and time consuming. One of the most repeated tasks is the creation of pivot tables, which are curated and annotated by lexicographers by harvesting different portions of digitized data. An example table related to questionnaires linked to the colors subject is presented in Figure 2.

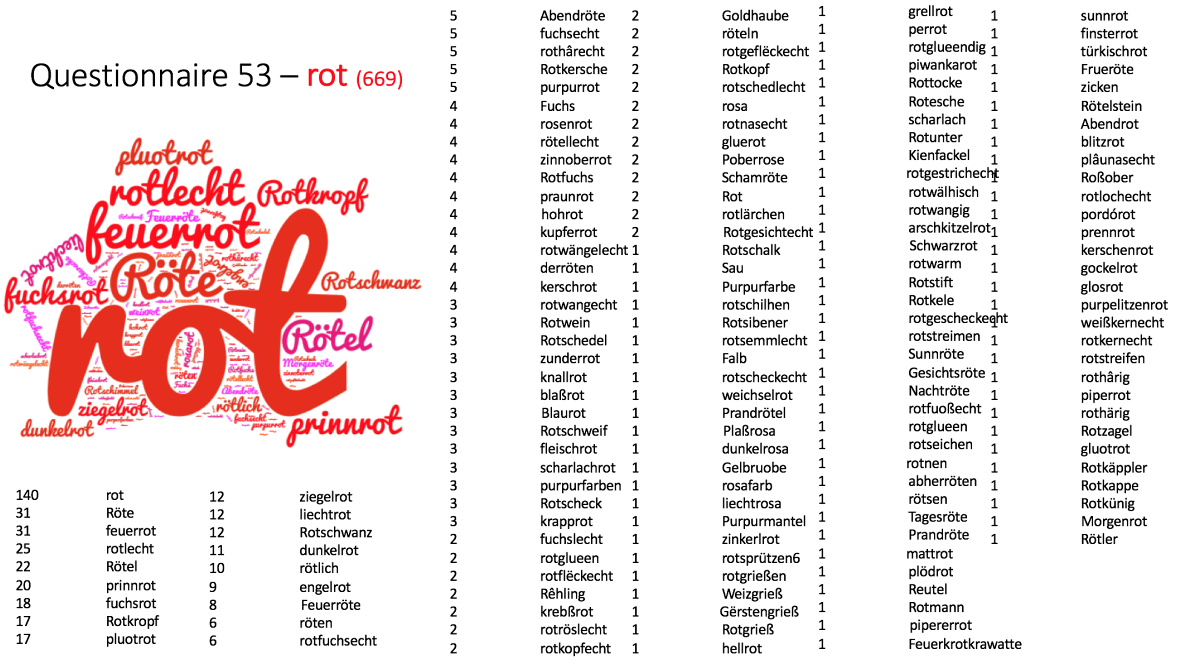
In this table the different lemmas pertaining to color-related questionnaires are represented. The lexicographers add information on top of the raw data like in columns D (Questionnaire), E (Question Topic) and F (Question specifier). They usually perform textual searches on the table and cross the results with entries related to the same lexical root or other textual features.



1	NR	NR/KT1	NR/KT2	Questionnair	Question T	Question s	Naming Referei	HL	NL	Gramma	Gramma	LT1	LT2	LT3	Meaning	BD/LT1	BD/LT2
2	53B7: schwärzlich	53B7: schwärzlich		COLOURS	BLACK	adjective		schwarzecht:	schwarzlecl	ADJECTIVE	simple	schwårtzlad					
3	53B7: schwärzlich			COLOURS	BLACK	adjective		schwarzecht:	schwarzlecl	ADJECTIVE	simple	schwarz!al¡at				schwärzlich	
4	53B7: schwärzlich			COLOURS	BLACK	adjective		schwarzecht:	schwarzlecl	ADJECTIVE	simple	schwarzl¡at				schwärzlich	
5	53B7: schwärzlich			COLOURS	BLACK	adjective		schwarzecht:	schwarzlecl	ADJECTIVE	simple	>zbuErtsElat				schwärzlich	
6	53B7: schwärzlich			COLOURS	BLACK	adjective		schwarzecht:	schwarzlecl	ADJECTIVE	simple	schwo¡arzlacht				schwärzlich	
7	53B7: schwärzlich			COLOURS	BLACK	adjective		schwarzecht:	schwarzlecl	ADJECTIVE	simple	>sw;ortsleXt				schwarzlecht, schwärzlich	
8	53B7: schwärzlich			COLOURS	BLACK	adjective		schwarzecht:	schwarzlecl	ADJECTIVE	simple	schwoa¡rzlat				schwärzlich	
9	53B7: schwärzlich			COLOURS	BLACK	adjective		schwarzecht:	schwarzlecl	ADJECTIVE	simple	>sw;oXtßlAd				schwärzlich	
10	53B7: schwärzlich			COLOURS	BLACK	adjective		schwarzecht:	schwarzlecl	ADJECTIVE	simple	schwǫazlad				schwärzlich	
11	53B7: schwärzlich			COLOURS	BLACK	adjective		schwarzecht:	schwarzlecl	ADJECTIVE	simple	schwoa¡rzlad				schwärzlich	
12	53B7: schwärzlich			COLOURS	BLACK	adjective		schwarzecht:	schwarzlecl	ADJECTIVE	simple	schwózl)at				schwärzlich	
13	53B7: schwärzlich			COLOURS	BLACK	adjective		schwarzecht:	schwarzlecl	ADJECTIVE	simple	schwoatzl))ad				schwärzlich	
14	53B7: schwärzlich			COLOURS	BLACK	adjective		schwarzecht:	schwarzlecl	ADJECTIVE	simple	>swoEtslEt					
15	53B7: schwärzlich			COLOURS	BLACK	adjective		schwarzecht:	schwarzlecl	ADJECTIVE	simple	schwarzlet					
16	53B7: schwärzlich			COLOURS	BLACK	adjective		schwarzecht:	schwarzlecl	ADJECTIVE	simple	schwoarzlat					
17	53B7: schwärzlich			COLOURS	BLACK	adjective		schwarzecht:	schwarzlecl	ADJECTIVE	simple	schwôa¡rzlåt				schwärzlich	
18	53B7: schwärzlich			COLOURS	BLACK	adjective		schwarzlecht	schwarzlich	ADJECTIVE	simple	schwårzlach				schwärzlich	
19	53B7: schwärzlich			COLOURS	BLACK	adjective		schwarzlecht	schwarzlich	ADJECTIVE	simple	>swåErzElich				schwärzlich	
20	53B7: schwärzlich			COLOURS	BLACK	adjective		schwerzlecht:2		ADJECTIVE	simple	schw.eschdslat				schwärzlich	
21	53B7: schwärzlich			COLOURS	BLACK	adjective		schwerzlecht:2		ADJECTIVE	simple	schwirzlad				schwärzlich	
22	53B7: schwärzlich			COLOURS	BLACK	adjective		schwerzlecht:2		ADJECTIVE	simple	schwi¡azlad					
23	53B7: schwärzlich			COLOURS	BLACK	adjective		schwerzlecht:2		ADJECTIVE	simple	schwiatzlath				schwärzlich	
24	53B7: schwärzlich			COLOURS	BLACK	adjective		schwerzlecht:2		ADJECTIVE	simple	schwiazzlad					
25	53B7: schwärzlich			COLOURS	BLACK	adjective		schwerzlich:2		ADJECTIVE	simple	schwärzli				schwärzlich	
26	53B7: schwärzlich			COLOURS	BLACK	adjective		schwerzlich:2		ADJECTIVE	simple	schwiErzl.e				schwärzlich	
27	53B7: schwärzlich			COLOURS	BLACK	adjective		schwerzlich:2		ADJECTIVE	simple	schbeaz|licl	schbeaz|li|	am schbeaz|li|gstn [S,fl		schwärzlich	
28	53C2: rötlich, rotlecht			COLOURS	RED	adjective		r<ötellecht:2		ADJECTIVE	simple	rẹat!al¡at				rötlich (schwachrot schillernd)	
29	53C2: rötlich, rotlecht			COLOURS	RED	adjective		r<ötellecht:2		ADJECTIVE	simple	rEetELEt				rötlich	
30	53C2: rötlich, rotlecht			COLOURS	RED	adjective		r<ötellecht:2	rôtellecht:2	ADJECTIVE	simple	roatE^lat				rötlich	
31	53C2: rötlich, rotlecht			COLOURS	RED	adjective		r<ötellecht:2		ADJECTIVE	simple	r;eatElEt				rötlich, leichter ins rote schlagend	

Figure 2: A detail of the pivot table generated for data coming from questionnaire no. 53 (farbe)

After this is done, they run statistical analysis and compose histograms and other visual artifacts that depict the numerical distribution of the result sets. The process is tedious as it is barely computer-assisted and it can take up to several days of work to reach a meaningful result set. The process does not stop there and there is one final step in which the researchers have to compile these sets into a visual artifact. In Figure 3 a histogram in the shape of a word cloud is presented. On it, frequencies of lemmas pertaining to a textual search for the lexeme rot are depicted, giving an idea of the nature of the analyzed data. It can be clearly seen what are the most common terms (rot-red) or (rte - redness), but also unexpected words as “weikernecht”.



140	rot	12	ziegelrot
31	Röte	12	liechtrot
31	feuerrot	12	Rotschwanz
25	rotlecht	11	dunkelrot
22	Rötel	10	rötlich
20	prinnrot	9	engelrot
18	fuchsrot	8	Feuerröte
17	Rotkropf	6	röten
17	pluotrot	6	rotfuchsecht

5	Abendröte	2	Goldhaube	1	grellrot	1	sunnrot
5	fuchsecht	2	röteln	1	perrot	1	finsterrot
5	rothârecht	2	rotgeflëckecht	1	rotglueendig	1	türkischrot
5	Rotkersche	2	Rotkopf	1	piwankarot	1	Frueröte
5	purpurrot	2	rotschedlecht	1	Rottocke	1	zicken
4	Fuchs	2	rosa	1	Rotesche	1	Rötelstein
4	rosenrot	2	rotnasecht	1	scharlach	1	Abendrot
4	rötellecht	2	gluerot	1	Rotunter	1	blitzrot
4	zinnoberrot	2	Poberrose	1	Kienfackel	1	plâunasecht
4	Rotfuchs	2	Schamröte	1	rotgestrichecht	1	Roßober
4	praunrot	2	Rot	1	rotwälhisch	1	rotlochecht
4	hohrot	2	rotlärchen	1	rotwangig	1	pordórot
4	kupferrot	2	Rotgesichtecht	1	arschkitzelrot	1	prennrot
4	rotwängelecht	1	Rotschalk	1	Schwarzrot	1	kerschenrot
4	derröten	1	Sau	1	rotwarm	1	gockelrot
4	kerschrot	1	Purpurfarbe	1	Rotstift	1	glosrot
3	rotwangecht	1	rotschilhen	1	Rotkele	1	purpelitzenrot
3	Rotwein	1	Rotsibener	1	rotgescheckecht	1	weißkernecht
3	Rotschedel	1	rotsemmlecht	1	rotstreimen	1	rotkernecht
3	zunderrot	1	Falb	1	Sunnröte	1	rotstreifen
3	knallrot	1	rotscheckecht	1	Gesichtsröte	1	rothârig
3	blaßrot	1	weichselrot	1	Nachtröte	1	piperrot
3	Blaurot	1	Prandrötel	1	rotfuoßecht	1	rothärig
3	Rotschweif	1	Plaßrosa	1	rotglueen	1	Rotzagel
3	fleischrot	1	dunkelrosa	1	rotseichen	1	gluotrot
3	scharlachrot	1	Gelbruobe	1	rotnen	1	Rotkäppler
3	purpurfarben	1	rosafarb	1	abherröten	1	Rotkappe
3	Rotscheck	1	liechtrosa	1	rötsen	1	Rotkünig
3	krapprot	1	Purpurmantel	1	Tagesröte	1	Morgenrot
2	fuchslecht	1	zinkerlrot	1	Prandröte	1	Rötler
2	rotglueen	1	rotsprützen6	1	mattrot		
2	rotflëckecht	1	rotgrießen	1	plödrot		
2	Rêhling	1	Weizgrieß	1	Reutel		
2	krebßrot	1	Gërstengrieß	1	Rotmann		
2	rotröslecht	1	Rotgrieß	1	pipererrot		
2	rotkopfecht	1	hellrot	1	Feuerkrotkrawatte		

Figure 3: A manually-generated word cloud representing the result set for the textual search “rot”

By understanding the process of creation of these tables, the data visualizer was able to create examples that replicated the aforementioned work flow by means of applying different computer-assisted methods. This greatly improved the times involved in the creation of visual artifacts like the one in the image, and fostered the prompt arrival at meaningful conclusions during the live coding sessions that took place in the following days of the STSM.

5 Live sessions

During the prepared live coding/analysis sessions the data visualizer showed examples to the team of lexicography experts in order to foster knowledge extraction out of the recently imported data sets and with the aim of acquiring insight on them. During these sessions, the data visualizer created ad-hoc interactive examples according to the flow of the ongoing discussion. The tools used were 1. The D3.js visualization library[1] and 2. Kibana[3]. In the following lines the most relevant examples are presented:

5.1 Meaning/Lemma Network

During the first session, the discussion led to the analysis of semantic groups of lemmas, as it was one of the missing key features that the existing prototype lacked. In this example, the most common meanings in the database were analyzed by means of a network graph, which shown how different lemmas were related to the same meaning. This type of network analysis is of proven efficiency in study of dialectal data and linguistic accidents in general[6], [7]. A screen capture of the results is presented in Figure 4:

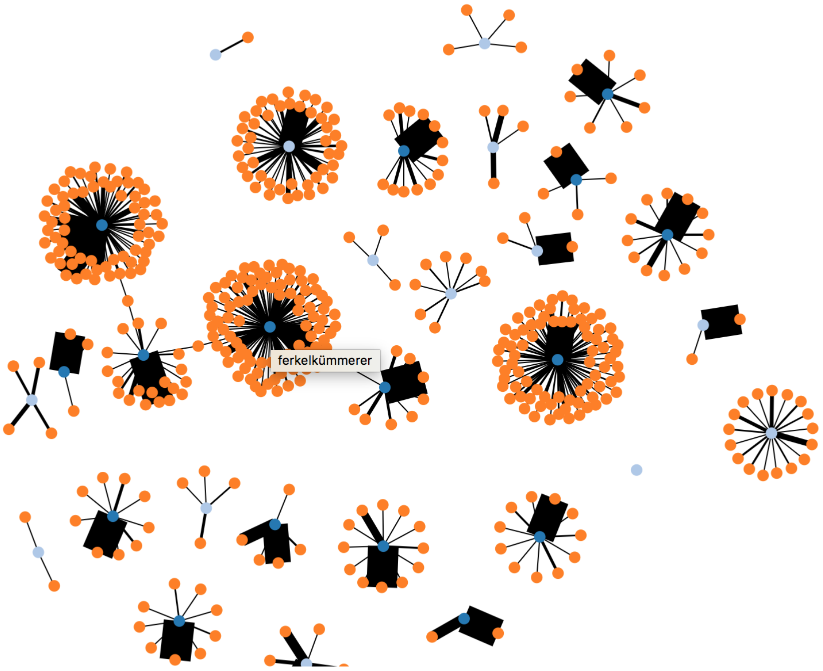


Figure 4: Comparison of the same lemma entry in the TUSTEP (left) and TEI (right) format representations

In this representation each node of the graph can be one of three different categories, adopting a specific color for each case:

1. Meaning, represented with dark blue color.

2. Lemma, represented with orange color.

3. Both, represented with light blue color.

The lines between the nodes represent a connection between a lemma and a specific meaning, denoting that an entry relating the two exists in the dictionary database. The thickness of the edge encodes the number of entries found, so the thicker the line the bigger the number of different sources is for a certain meaning-lemma pair.

As a result of the application of this technique, different lexical groups could be visualized and explored. As it was expected, each unit is formed by a central meaning connected to several different lemmas. In the figure we see the label for the meaning unit of *ferkelkmmerer*, which is the first piglet of a litter. In the example it is also shown how this lexical group is connected to others through lemmas connecting in turn with other meaning. In the code example it could be verified that the lemma *gaumen* connects the semantic units of *ferkelkmmerer* and *schauen*. This shed some light on the possibilities and applications of new versions of the prototype that treated with this kind of network analysis in ways that are useful for the lexicographers' research tasks.

5.2 Dictionary Data Metrics

Another tool that was used to analyze other aspects of the historical dictionary data was Kibana. This tool works in parallel with the textual search engine and displays metrics and other useful information that help stakeholders to form a mental image of the nature of the data, as it was explained in section 4.

In our approach we were able to provide visual representations of the different result sets related to textual queries performed by the lexicographers. In opposition to the manual methods usually employed by the lexicographers, this methodology greatly reduced the times (from hours to just minutes) needed to reach a visual representation. In Figure 5 a result set for the most common meanings is presented. Once the system was set-up and running, it was just a matter of seconds to arrive to the results depicted in the image.

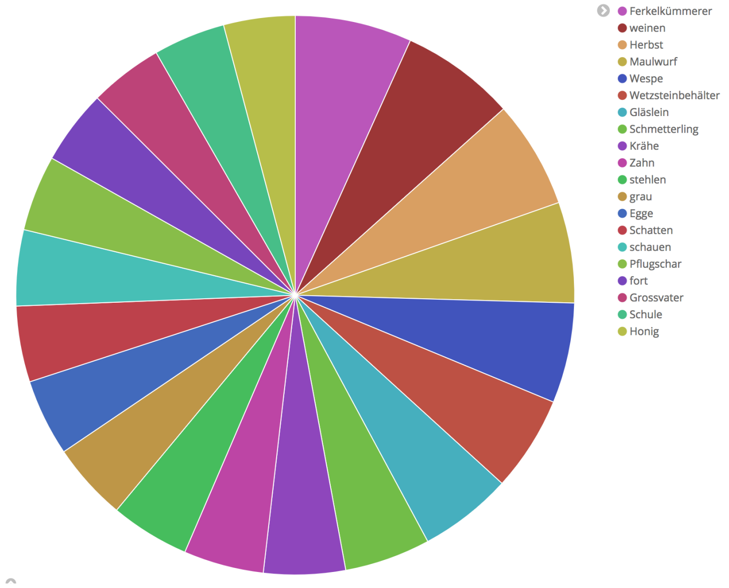
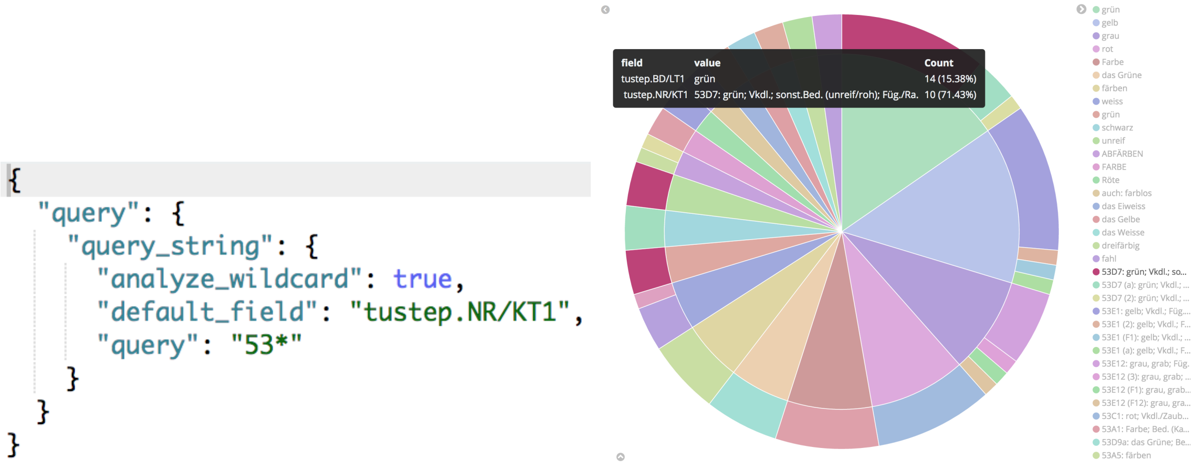


Figure 5: Histogram depicting the statistical distribution of the 10 most popular meanings in the data set

Not only it was possible to display general metrics represented in visual, easy-to-understand ways in short times but also we could perform textual searches that mimicked those of the lexicographers. For instance, we were able to replicate some of the examples we introduced in previous sections by employing full-text queries on the data set. In Figure 6 we illustrate the query, which retrieves all the entries related to questionnaires starting by “53” (related to color). With this result set we create a histogram representing the different distributions for the two represented dimensions (lemma and questionnaire number).



```
{
  "query": {
    "query_string": {
      "analyze_wildcard": true,
      "default_field": "tustep.NR/KT1",
      "query": "53*"
    }
  }
}
```

field	value	Count
tustep.BD/LT1	grün	14 (15.38%)
tustep.NR/KT1	53D7: grün; Vkdl.; sonst.Bed. (unreif/roh); Füg./Ra.	10 (71.43%)

Figure 6: Full-text query that retrieves all the entries for questionnaires starting by 53 (left). On the right, the visual representation of the result set

In the visualization it can be seen what the most popular lemma is (grn-green) and what questionnaires these entries belong to (53D7, 53D7(a) and 53D7(b)).

6 User Stories

During the course of the STSM, the different stakeholders were able to identify and refine several user stories that will be converted into software requirements, which in turn will be coded as features in the next version of the visual exploration prototype currently being built. Of special importance are those related to the lexicographers, which are listed below in the following listing:

As a humanist I want...

- to represent time stamps of my result set because the temporal evolution of an entry is relevant to my research.
- to know how words referring to the same concept change over time in certain area because the information reveals in cultural context how conservative a dialect region is.
- to run statistical analyses (i.e. frequencies, distributions, deviance) on the result sets because I want to have a numerical overview of the results.
- to represent uncertainty in time stamps (i.e. 19xx; 19x[25-27]; (1936+1978), persons, locations of my result set and correlate entries according to other dimensions because completing the gaps in terms of missing information in the data and filling them.
- to select and extract terms or parts of terms of my results set, group them accordingly and display them on a map and see how they evolved in time because they show the diversity in Bavarian dialects.

- to show the evolution of terms of my result set on a map because pronunciation features are indicative of the historical dialect areas.
- perform a search on mainlemma, lautung, sense and group search results by their pronunciation because to find out if similiar terms have similar/different meanings and how these relate to time and/or place and/or person in order to discover/validate dialect areas.

Although some of the most important of these user stories were treated during the STSM, others could not be addressed given the limited length (5 working days) of it and thus they will be studied in the following months when the first, most basic ones have been completed.

7 Conclusions

The STSM presented in the previous sections supposed a great achievement for the writer of this document in the understanding of some of the most basic work flows in modern lexicography. The possibility to stay at the Academy of Sciences working along many of the experts has been key to further develop my cognition and interpretation of the work related to the dictionary assembly and curation processes, as well as all the parallel work related to the them. By further examining the new data formats and user stories, it will be possible to connect historical dictionaries data with other data sets coming from other European languages and institutes, in an attempt to visually represent all the linguistic richness within the Union.

References

[1] D3.js home page.

[2] Elasticsearch home page.

[3] Kibana home page.

[4] A. Benito, A. G. Losada, R. Therón, A. Dorn, M. Seltmann, and E. Wandl-Vogt. A spatio-temporal visual analysis tool for historical dictionaries. In *Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality*, TEEM '16, pages 985–990, New York, NY, USA, 2016. ACM.

[5] O. A. der Wissenschaften (AW). Biodiversity and linguistic diversity project, 2016.

[6] W. J. Heeringa. *Measuring dialect pronunciation differences using Levenshtein distance.* PhD thesis, Citeseer, 2004.

[7] T. Mayer, J.-M. List, A. Terhalle, and M. Urban. An interactive visualization of crosslinguistic colexification patterns. *09: 00–10: 30–Morning Session, Part I 09: 00–09: 10–Introduction 09: 10–09: 40 Thomas Mayer, Johann-Mattis List, Anselm Terhalle and Matthias Urban, An Interactive Visualization of Crosslinguistic Colexification Patterns*, 11(15):1.