A New LMF Schema Application by Example of an Austrian Lexicon Applied to the Historical Corpus of the Writer Hugo von Hofmannsthal

Armin Hoenen and Franziska Mader

Abstract

In this paper, that goes along with the release of an Austrian lemma list for NLP applications, the creation and representation of a digital dialect lemma list from existing internet sources and books is presented. The creation procedure can serve as a role-model for similar projects on other dialects and points to a new cost saving way to produce NLP resources by use of the internet in a similar way to human-based-computation. Dialect lexica can facilitate NLP and improve POS-tagging for German language ressources in general. The representation standard used is LMF. It will be demonstrated, how this lemma list can be used as a tool in literature science, linguistics and computational linguistics. Especially the critical edition of Hugo von Hofmannsthal is a well-suited corpus for the aforementioned research fields and the inspiration to build this tool.

1 Introduction

In NLP for various tasks such as POS-tagging, stemming, information retrieval and so forth lexical resources are employed. In order to study corpora with dialectal components, the lexicon must contain the dialect words. For the digital representation of lexica, various standards have emerged throughout the last decades. The ISO standard of LMF (ISO24613, 2005), Lexical Mark-up Framework, is one of the most versatile platforms for lexicon representation and it integrates several state-of-the-art features. In its beginning a UML specification, the website `http://www.lexicalmarkupframework.org/` provides an XML DTD [1] with one base file and several extensions. For the linguistic features LMF uses another ISO standard encapsulated into a "feat" tag, the Data Category Registry[2].

In the next section we propose an LMF model. In section 3 we present the application framework for which the lexicon has been designed, followed by a recapitulation of the peculiarities of Austrian German and the description of the reusable workflow of a dialect lexicon creation procedure. The corpus investigated in section 6 contains the works of Hugo von Hofmannsthal an Austrian writer who lived between 1874 and 1929 thus subject to the effects of two orthographic conferences. It encompasses the critical edition volumes 6, 7, 16, 17, 19, 21, 22, 25-1, 27, 33 and 34 and includes amongst others drama, poems, essays and narratives. Section 8 concludes with a brief summary and an outlook.

2 The LMF Model

Various standards for the digital representation of lexica have emerged. Detailed descriptions of the developments can be found in (Budin et al., 2012). Francopoulo et al. (2006) give an insight

[1] `http://www.tagmatica.fr/lmf/DTD_LMF_REV_16.dtd`

[2] `http://www.isocat.org/`

into the emergence of the ISO standard of LMF (Francopoulo et al., 2006), that integrated various former projects. LMF as an ISO standard was chosen as the primary representation format, but it is planned to offer TEI and RDF as alternative data exchange formats. Based on the two ISO standards of LMF and Data Category Registry (DCR) "that is maintained as a global resource by ISO TC37" (Francopoulo et al., 2006), a new representation model for lexica has been worked out, which is being applied to a number of historical corpora within the LOEWE initiative of the state of Hesse[3]. The following representation based on these standards was developed for the integrative use in the module called Lexicon Browser within the humanities computing platform of the eHumanities Desktop (Mehler et al., 2009) described in section 3.

2.1 The Header

The header specification follows the LMF proposal quite strictly and incorporates the structure in a minimal way. Language and encoding are represented.

Listing 1: The LMF header for a Lexical Resource.

```
<GlobalInformation>
 <feat att="languageCoding" val="ISO␣639-3" />
</GlobalInformation>
<Lexicon>
 <feat att="language" val="deu" />
 ...
```

2.2 Lexical Entries

However, the representation of Lexical Entries can become complex. In the following example, a dialectal/historical spelling variant is encoded by means of a *LexicalEntry* featuring *FormRepresentations* of its *WordForms*.

Listing 2: An example word form.

```
<LexicalEntry id="18339941">
 <Lemma>
  <feat att="type" val="dialectal" />
  <feat att="dialect" val="Austrian" />
  <feat att="label" val="Abschnitzel" />
  <feat att="description" val="abgeschnittenes␣kleines␣
      Stueck" />
  <feat att="part␣of␣speech" val="noun" />
  <feat att="gender" val="n" />
 </Lemma>
 <WordForm>
  <FormRepresentation>
   <feat att="id" val="18339942" />
   <feat att="label" val="Abschnitzel" />
   <feat att="case" val="nominativeCase" />
   <feat att="number" val="sg" />
  </FormRepresentation>
  <FormRepresentation>
   <feat att="id" val="18339943" />
   <feat att="label" val="Abschnitzl" />
   <feat att="case" val="nominativeCase" />
   <feat att="number" val="sg" />
  </FormRepresentation>
 </WordForm>
 <WordForm>
  <FormRepresentation>
   <feat att="id" val="18339944" />
   <feat att="label" val="Abschnitzels" />
   <feat att="case" val="genitiveCase" />
   <feat att="number" val="sg" />
  </FormRepresentation>
 </WordForm>
 ...
```

Each lexical entry by definition must have a *Lemma*, which carries the attributes of type, id, name, description and part of speech by default plus additional features that are not subject to change within the inflectional paradigm of the present part of speech. These additional grammatical attributes are features of the *WordForm*. For a noun, for instance, gender is a feature of the *Lemma*, as it never changes regardless of case, while case itself is a feature of the *WordForm* which is therefore not present for the *Lemma* [4]. Of course these restrictions are language specific and they must be specified by the user upon input. Yet another feature of the *Lemma* is the type, which in our system can have the value "dialec-

[3] http://www.digital-humanities-hessen.de/

[4] If a wordform happens to have a different part of speech and therefore a conflicting value in any of the parent-lemma's default set, the word form's feature is spelled out and overwrites the lemma's feature.

tal"; then, the dialect's name is specified by the next feature. The lemma's id corresponds to the id of the lexical entry as a whole, that is the lexicon is sorted by lemmata as the upmost hierarchical layer. If an entry has different spellings, as is often true for historical or dialectal word forms, those are encapsulated in a *FormRepresentation*, otherwise containing the same information each as have non-variant word forms. All *FormRepresentations* make up for one word form. It is exactly this representation, that is used to display dialectal and variant writings. *Synsets* are used to group synonymous semantics or senses for dialectal items and their standard counterpart if existant. The representation format was incorporated in LMF from WordNet. Additional export formats in the near future will include RDF and TEI.

3 The Application Framework

Any lexicon, which uses the above described LMF as an input/output format can be managed within the eHumanities Desktop, a humanities computing platform accessible through the browser. The eHumanities Desktop is a web-interface allowing users to share and organize resources but also to analyse them. Once uploaded into the Lexicon Browser via LMF, the user interface makes a lexicon browsable, performs search operations and obtains statistics connected with each single entry. As can be seen on Figure 1, the interface shows for instance word forms connected to a query and provides respectively grammatical information in a human readable way. Additionally, it displays the information graphically in a network. The lexicon can be connected with a text. If so, all occurrences of the word forms are linked with the text, frequency distributions and collocation statistics are available through another module within the eHumanities Desktop called *Historical Semantics Corpus Management*, see (Jussen et al., 2007; Mehler et al., 2011).

The user can annotate, reannotate and perform an online reindexation in order to keep the (statistical) information up to date. In (Gleim et al., 2012) the application framework of lexica and corpora management and its architecture are being described in greater detail.

4 Austrian German

German is a so-called pluricentric language (Clyne, 1992), that is, there is more than one center from which various standardization processes spread, leading to a mosaic of different partly overlapping substandards, varieties and dialects. Additionally, a plurality of countries with German as a national language exist.

One of those countries is Austria. Austrian German has as many as three different neighbouring non-germanic language families (Slavic, Finno-Ugric and Romance) plus some additional sources for calques and loans (like Yiddish or Rotwelsh) (cmp. (Beyerl et al., 2009), (Wiesinger, 1990)). Research on Austrian began at least as early as 1774 under the empress Maria Theresia under whom the abbot Johann Ignaz Felbiger created a schoolbook with first lists of Austrian terms (Back et al., 2009). The orthography of German in Austria was administered in Vienna while throughout the 19th century Prussia and other emerging German regions kept defining their own standards. In order to resolve differences within the German speaking lands two orthographic conferences (1876 and 1901) were held.

Research encompassing the Austrian variety of German in former times has led to the production of various printed lexica, the most important of which continues to be used in Austrian schools (Back et al., 2009). There is also an EU protocol of some 30 Austrian terms with their counterparts in Standard German (Markhardt, 2005). On the internet, various sites with dialectal content can be found in guestbooks, forums and chats ((Bashaikin, 2005, 444)) and a wikipedia for the Bavarian dialectgroup exists(`http://bar.`

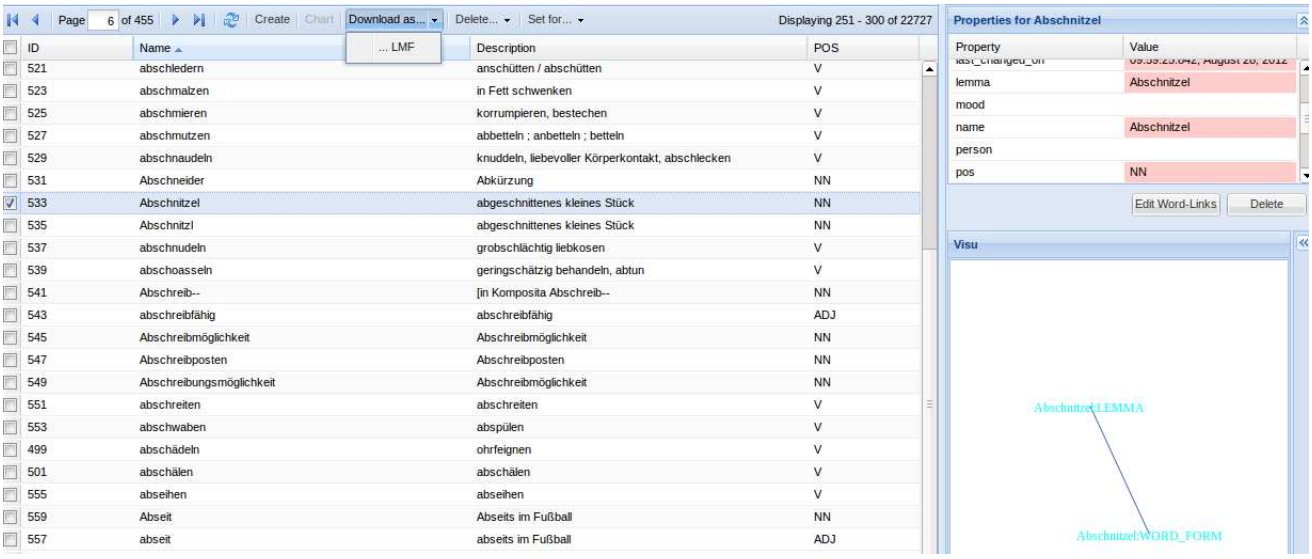

Figure 1: The Lexicon Browser.

wikipedia.org/). Still, to the best knowledge of the authors there is no publicly available free digital annotated dialect lexicon or word list. The ICLTT[5] offers the "Wörterbuch der bairischen Mundarten in Österreich" on a commercial basis. Linguistically, Austrian dialects belong to the Bavarian dialect continuum. Different subvarieties are attested (see for instance (Rowley, 1990), (Wiesinger, 1990)). This adds an element of complexity to the lexicon structure. Concerning the orthography, Auburger (2011) notes, that there is no widely accepted standard yet for the written manifestation of the Bavarian dialects. Nevertheless, a Wikipedia in this variety exists. In written language the following features are widely applied non-phonological ones distinguishing the dialect from the standard((Wiesinger, 1990)):

1. lexicon [6]
2. diminutives [7]
3. 2nd plural for verb forms [8]
4. gender differences [9]
5. differences in the use of prepositions
6. word formation with the 'to be' auxilliary for all verbs of motion [10].
7. for additional features see (Wiesinger, 1990)

[5]Institut für Corpuslinguistik und Texttechnologie Austrian Academy of Sciences

[6]Certain words like "Schmäh" (nonsense) or terms from cuisine like "Zibebn"(raisins).

[7]While Standard German uses either of the suffixes "-lein" or "-chen" to form a diminutive (Meibauer, 2007). Austrian German uses a reduced form of the first of those "-l". Phonological rules of assimilation with this suffix differ (Mauser, 2005) and it is generally more productive in southern German varieties than in the standard. An example is "Land" (land) and the Austrian diminutive form "Landl" as opposed to the possible Standard German forms "Ländlein/Ländchen" and the Allemanic form "Ländle".

[8]The second plural is marked by an *s* distinguishing it from the first and third person plural in verbal inflection, while in Standard German and Dutch the second person plural forms are not marked. "Ihr gebts des dem Hansl." as opposed to "Ihr gebt das dem Hansi/Hänslein/Hänschen." (You will give this to Hans.)

[9]"das Teller" (the plate) as opposed to "der Teller"

[10]In Standard German mostly with 'to have': "ihr seids da gesessn" (you've sat there) vs. "ihr habt da gesessn"

5 Lexicon Creation

In order to improve the performance of our postagging for the works of the Austrian writer Hugo von Hofmannsthal (1874-1929), it was decided to create a digital lexicon for Austrian German. Although the critical edition in its apparatus explains austriacisms once they appear and contains a glossary in volume XXXIV, a more general lexicon would be applicable to comparable texts by other authors as well. The glossary was included along with single translations while the internet was taken as the source for the majority of entries. Certain sites provide lexica for Austrian. The plupart of these are created and maintained by laymen and some are cooperative sites, where each entry emerges from a blogger, who adds it. As (Geyer, 2005) points out:

"Großlandschaftswörterbücher wie das Wörterbuchs der bairischen Mundarten in Österreich (WBÖ) haben eine lange Sammelphase und eine lange Publikationsdauer"[11]

(Geyer, 2005, 195)

Geyer lists the timerange of the emergence of this lexicon. It accounts for 107 years (1913-2020). The main sources are just like the digital ones informants, that is laymen, who have in this case answered questionnaires. The procedure being applied is making use of the internet in much the same way as is human-based-computation. The people who have created the lexicon sites for the public enabled the project to create a lexical resource without the time consuming process of constructing questionnaires. This shortens the process of producing an annotated NLP dialect lemma list significantly. However, the core of the resource having emerged without scientific control, an estimate on how comprehensive or how biased towards one of the subvarieties the resource may be is hardly possible without further anlyses.[12] The successfull application of the lemma list will be demonstrated in Chapter X.

With several substandards and differing dialectal orthography, taking a merger of all of the sites seems to be more reliable than, for instance, taking only the biggest one. Yet, a consequence of merging is some additional effort in the creation of the ressource. The following sites were taken as a basis for the creation:

`oesterreichisch.net`, `oewb.retti.info`, `ostarrichi.org`, `unsere_Sprache.at`, `de.wikipedia.org/wiki/Liste_von_Austriazismen`, `openthesaurus.de/synset/variation/at`, `sistlau.blogspot.de/`, `german.about.com/od/vocabulary/a/Austrian.htm` and `das-oesterreichische-deutsch.at`.[13]

The EU Protocol 10 was also taken into account as well as entries from a dialect guide published by Beyerl et al. (2009). After having identified these resources, the steps in lexicon creation that followed were:

1. downloading the entries from the internet
2. unifying the format
3. merging the entries
4. resorting and deleting duplicate entries
5. removing Standard German entries without the loss of false friends
6. detecting and relating synonyms
7. annotating pos-tags
8. test the application of the lexicon

[11] Large-area-lexica like the lexicon of the Bavarian dialects in Austria (WBÖ) have a long collection period and a long publication phase.

[12] An analysis of the contained writing variants in the light of non standardised orthography seeking parallels to historical language phenomena is envisaged for subsequent research.

[13] All March 2012.(?pruefen)

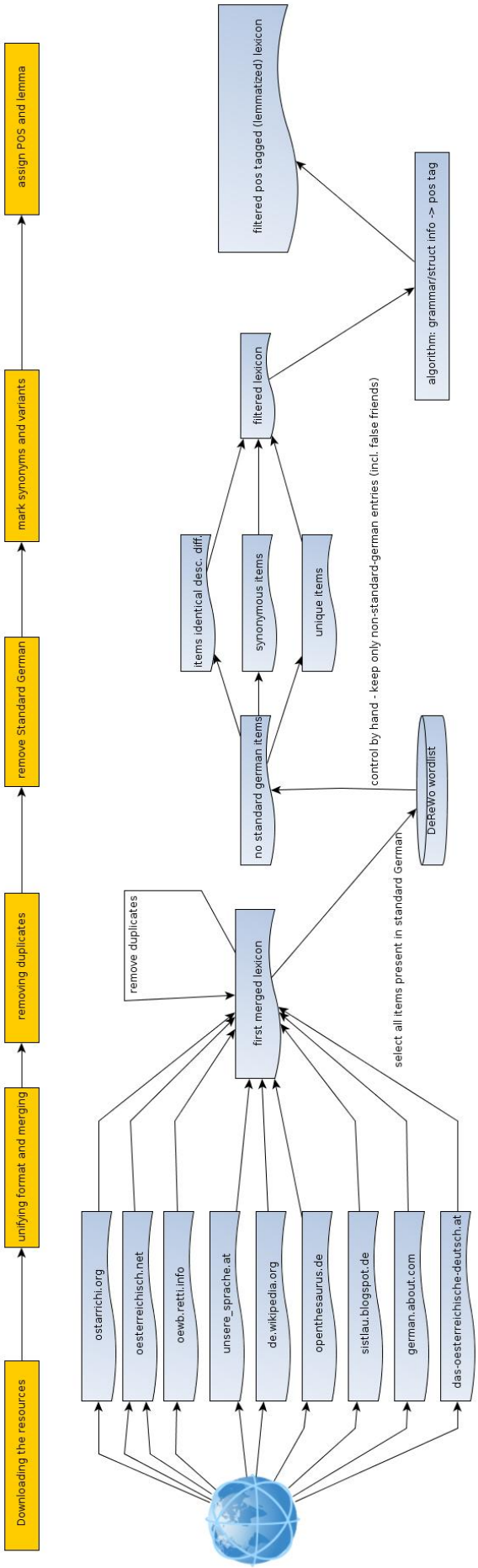

Figure 2: The workflow for the lexicon creation.

Each step had an input and an output as shown in the workflow diagramm, Figure 2. In order to ensure correctness, a lot of manual labour was involved for control of the output of each step. The workflow is a general procedure that can be applied to any such task. Entries were first downloaded through a script in the Java programming language followed by a unification of format. The chosen format is held in a table-layout, featuring columns and rows, where each row is a new word form, while each column has a new information type. The first column contains the Austrian word, the second column has additional grammar information if present, the third column has the meaning or description, the fourth column features additional information and the last column contains the sourcesite. Later on, two additional columns for POS tag and lemma were filled.

5.1 Removing Standard German entries without losing false friends

In order to avoid the capture of Standard German words, tokens had to be removed, if in both meaning and spelling they were congruent. As has been shown several times, there is no objective criterion to draw a clear line between the notion of language and dialect. For the authors of dialectal lexica, this is a challenge. Certain words will clearly be borderline cases. Some of the judgements might even depend on the command of standard and dialect of the lexicon author. For the sake of consistency and comprehensiveness, every word can be accepted for input, so as to certainly capture all the core items. However, for the reasons mentioned above[14], in order to be more restrictive, a check for entries overlapping with Standard German was performed. The list of the most frequent 100 000 German words as published by the IDS Mannheim [15] has been intersected with the lexicon. The widely known items were removed. Many such words were colloquial or curse words, a category barely written and therefore more easily perceived as dialectal (considering that the formula ”that which you see written is the standard, that which you hear is the dialect” is for most people easy to understand and memorize and at the same time a sufficient explanation for the dichotomy of language and dialect). If a Standard German term would have a completely different meaning in Austrian German, commonly labeled ”false friend”, it would be kept. A more subtle case for the decision on tokens were entries, which had a Standard German equivalent with the exact same meaning, but which were affected by minor sound alternations. In principle, the aim of the lexicon was to capture those elements that are not present in dictionaries of the standard and which are not necessarily detected as variants by established measures like the Levenshtein distance (Levenshtein, 1965). Hence, the tokens with cognates were only accepted, if they appeared in the source lexica and if the number of phonological differences (or their degree) rendered the token incomprehensible if appearing in a standard context, for the decision upon which native speaker intuition was used as the benchmark.

5.2 Synonyms and Variants

Another step in the creation of the lexicon was the treatment of synonyms and variants. Synonyms were treated as interconnected itemsets with a set of synonym relations to the ids of other tokens and tokens with different senses were duplicated and displayed as non- connected. Table 1 and Table 2 display token relationships.

[14] For instance, an Austrian term, which is widely used in the standard and has no immediate alternative synonym there, should also not be counted as an Austriacism.

[15] `http://www.ids-mannheim.de/kl/derewo`

unique ID	token	translation	synonymset
9541	Hopertatsch	ungeschickter Mensch	[9541,9550,23160]
9550	Hoppadatschi	ungeschickte Person	[9541,9550,23160]
23160	Hirsch	ungeschickte Person	[9541,9550,23160]

Table 1: Synonyms and Variants

unique ID	token	translation
9558	hoppertatschert	überheblich/ungeschickt
9559	hoppertatschig	überheblich/ungeschickt
9558	hoppertatschert	überheblich
9559	hoppertatschert	ungeschickt
9560	hoppertatschig	überheblich
9561	hoppertatschig	ungeschickt

Table 2: Separating items - above: state of entries before manual separation; below - separated senses

5.3 POS-tagging and Lemmatizing

In order for the lexicon to become a useful NLP resource, basic POS (ADJ, NN, V, PART, ADV) were annotated. Luckily, German nouns are capitalized, so with very great certainty an entry being capitalized was a noun; verbs were more demanding, but obligatorily end in -n in the infinitive, which serves as lemma for German, so we applied this rule. Adjectives were most diverse and all of the automated pos assignments were controlled by hand after the automatic preprocessing. The automatic POS-assignment has been evaluated:

Lemmata were conversely easily annotated as a lexicon entry is already a lemma. At the end of this step the lemma list contained 19 479 tokens: 12 192 nouns, 3144 verbs, 1389 adjectives, 388 adverbs. 2061 entries contained at least one space character, thus at least 2 words (articles are not counted here; they had been separated already beforehand). They were typed as multi word unit (MWU). Additionally there were 305 items which were either particles, suffixes or had the possibility to be interpreted as more than one part-of-speech. They were manually annotated for part of speech.

wordclass	precision
(a) nouns	0.975
(b) verbs	0.96
(c) adjectives and other	0.65
mean ($\frac{a+b+c}{3}$)	0.93

Table 3: Automatic POS-tagging

5.4 Expansion

For the ca. $20,000$ obtained lemmas, according to inflectional paradigms given in (Wiesinger, 1990) an expansion scheme has been set up. For nouns, verbs and adjectives, we produced ca. $112,000$ wordforms, connected to their lemmata. An LMF version of the lexicon is available and part of the eLexicon for Austrian in the eHumanities Desktop (Mehler et al., 2009).

6 Detection

The lemma list can serve as an input to detect patterns of dialect usage throughout a text. In order not to capture items present in the standard language, the lexicon was separated into lexical and semantic austriacisms by again detecting the overlap with a big German lexicon used in the pos tagger published by (Waltinger, 2010). The lemma list was used for detection of austriacisms in a text. The detection was augmented by a simple matching of two idiosyncratic features of the Bavarian dialects:

- inflected auxilliaries in the second plural as listed by (Wiesinger, 1990) (derfts/dürfts, gehts, mögts, müssts, sollts, wollts immer mit ihr + habts seids tuts)
- endings characterising the Austrian diminutives without capturing false positives by a regular expression (Cl)
- relative clause entry sequence ”die wo” (Eroms, 2005)

7 Application of the Lexicon

7.1 Linguistics

In linguistics, code-switching refers to various patterns of ”the use of different languages in the same discourse” (Thomason, 2001). This applies not only to languages, but also to dialects (Niebaum and Macha, 2006, 9), where an additional layer of complexity is the degree of dialect usage interwoven with the standard. In Hofmannsthal, in the minority of his works, dialectal elements appear. If they do, a richness in gradation of the dialect can be seen, which actually appears in normal speech of dialect competent speakers all around the world. (Niebaum and Macha, 2006) reports of three typical types of dialect speakers identified by ++ Lausberg (1993), code-switchers, code-mixers and dialect speakers. In Hofmannsthals unfinished work ”Wiener Pantomime” (Viennese Pantomime) a text which was never actually published and is only attested in fragments, having not undergone any further writers or editorial processes, in the following three subsequent lines dialect is used in different ways.

der römische Kaiser: Jetzt wieder schlafen
ART roman emperor: Now again sleep-INF
gehen, Schmarr’n!
go-INF nonsense
The roman emperor: Going to sleep again now, nonsense!
die Schäferin:Ich möcht beim
ART shepherd-f1-Sg-NOM want-1st-Sg to-DEF
Calafati fahr’n!
Calafati(name) go-INF
The shepherd(f): I want to go to the Calafatti
der Herrnhuter: Was fallt Dir ein! Dir
ART Herrnhuter(name): what imagine(stem) PP-2-Sg-DAT imagine(preverb) PP-2-DAT
wer i’s zeig’n!
be-FUT-1-Sg 1-Sg-NOM’3-Sg-OBJ(reduced) show-INF
the Herrnhuter: What are you thinking! I’ll have the last laugh!

In the example the dialect has been applied in different degrees. The only word in the speech of the emperor which is clearly dialectal is his last word *Schmarr’n* (nonsense) a lexical austriacism. If we look at the personal pronouns and at the verbal forms, the shepherd and the Herrnhuter use different patterns. While the shepherd uses dialectal verbforms, but the standard language’s first person singular personal pronoun ”Ich”, the Herrnhuter uses both dialectal verbal inflection and the dialectal form of the first person singular pronoun (”i”). Thus they display different degrees of dialect application. Linguistic hypotheses could be for instance that:

- if dialectal pronouns are used, they are used for the entire class of pronouns, never only one

- if dialectal pronouns are used, dialectal verbforms must be used as well, but never the other way around (hierarchy)

With the lemma list especially in written corpora and most of all in the critical edition of Hofmannsthals texts, where the authors personal thoughts, correspondences and intentions on using the dialect are included in a comprehensive critical apparatus, these phenomena could be investigated in more detail by application of automatic detection through the lemma list presented here. A digital text together with a detection tool would in this case facilitate the process of data acquisition. Another example of dialectal usage in spoken English on the border of linguistics and literature science stems from (Gardner-Chloros, 2009) who finds for instance a potential narrative use of varieties. ”Sebba [a discourse participant] suggests that code-switching is used here

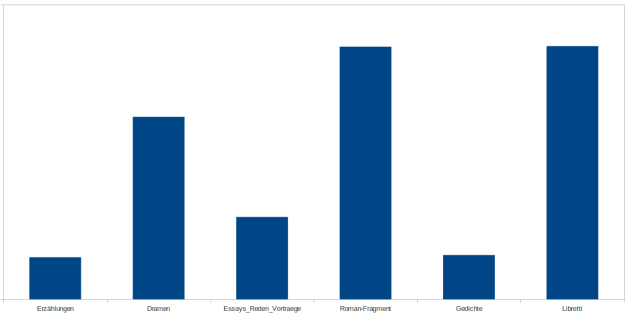

Figure 3: Part of the repetition typology.

to ”animate” the narrative by providing different ”voices” for the participants in the incident which is described.”(Gardner-Chloros, 2009, 3) She refers to code switching between Creole and standard English. This however is an example of unconscious usage, the next section will enlighten some aspects of conscious usage of dialectal components by an author for literary purposes.

7.2 Literature Science

With the aid of the dictionary for ”Austriazismen” digitized texts by Hugo von Hofmannsthal can be examined with respect to the appearance of Austrian terms. Via a quantitative analysis of the distribution of dialectal words statistics can reveal in which genre or volume of the critical Hofmannsthal edition austriacisms accumulate. Austrian words are not circumscribed to the lexical level but constitute diminutives and certain forms of inflection as well. Fig. 3 illustrates that besides the ”Roman-Fragment” especially Hofmannsthal’s scenic works (Dramen, Libretti) exhibit an ostentatious cluster of dialectal terms. An ostensive example is ”Der Schwierige” because explicit Austrian terms as well as correspondent grammatical forms appear in this play (Mauser, 1982, 115). The critical edition offers a list of foreign terms and their meaning in the appendix. Besides the Austrian words the list contains French and English terms as well and aggravates the differentiation of austriacisms. In his article about ”Der Schwierige” Wolfgang Mauser ascribes the application of the existing austriacisms to the accentuation of Austrian traditions (Mauser, 1982, 115): the protagonists are members of the Austrian nobility and thus prefer an exalted lifestyle. The location of the story line - Austria - plus the clientele of the play are considered as the motivation for the application of the correspondent dialect. Apparently the vernacular shall refer to a traditional consciousness and patriotism of the figures (Mauser, 1982, 115). The used dialectal forms not only describe specific Austrian customs or local dishes which are unknown to foreigners but are common terms such as ”schurigeln” (to bedevil) or ”tentieren” (to intent). Thus austriacisms are not only utilized in cases when no Standard German term is available but are applied instead of the High German. Furthermore Mauser claims that Hofmannsthal systematically exaggerates the application of Austrian terms and thus generates humour (Mauser, 1982, 115). The conclusion for the usage of the Austrian dialect would thus be to ironicise ancient Austrian traditions of the nobility.

Another starting-point regarding the analysis of the Austrian dialect as a stylistic device is the social context of figures. In 1919 Hofmannthal himself wrote in the magazine ”Die Theater- und Musikwoche” about his libretto ”Die Frau ohne Schatten”:

”Ich wollte das Ganze als Volksstück, mit bescheidener begleitender Musik, machen, zwei Welten gegeneinanderstehend, die Figuren der unteren Sphären im Dialekt.” (Hofmannsthal, 1998, 236).

The territorial hierarchy of ”high” and ”low” implies the difference in social status of the protagonists: it is about a royal couple on the one side and a dyer and his wife on the other side.

Whether the missing dialect in the speech of the sovereign couple can be ascribed to their exalted educational background must be left open.

Less obvious is the application of austriacisms in Hofmannsthals opera "Der Rosenkavalier". Although social hierarchies are illustrated via varying speech levels the boundaries between dialectal forms and High German are vague. The maid Mariandl speaks in the vernacular but her mistress and other aristocratic figures also include dialectal terms and phrases in their speeches (code-mixing). Hofmannsthal himself describes the setting of the opera - Vienna - as a city where social differences are mirrored in the manner of speaking:

"[...] dieses Wien von 1740, eine ganze Stadt mit ihren Ständen, die sich gegeneinander abheben und miteinander mischen, mit ihrem Zeremoniell, ihrer sozialen Stufung, ihrer Sprechweise oder vielmehr ihren nach Ständen verschiedenen Sprechweisen [...]" (Hofmannsthal, 1986, 549). The Austrian historian Adam Wandruszka emphasises the dialectal cadence in the "Rosenkavalier". He points out that this linguistic characteristic could be ascribed to a journal which was published a short time before Hofmannsthal began writing his libretto (Wandruszka, 1967, 562). This journal was written by the controller of Maria Theresia's household and describes the life at court in the contemporary Viennese dialect (Wandruszka, 1967, 562). Wandruszka hypothesises that Hofmannsthal was familiar with these texts (Wandruszka, 1967, 562).

Besides the authorized texts of Hofmannsthal his literary legacy contains a number of fragments that are published in the critical edition of the Freies Deutsches Hochstift in Frankfurt/Main.

The unfinished piece "Wiener Pantomime" introduces a traditional Viennese character, "den lieben August" as the protagonist. His speech is distinctive Austrian concerning the lexical, grammatical and syntactic level. In contrast to August other figures, for example the nymphs, are explicitly supposed to speak High German the way "children recite wishes" (Hofmannsthal, 2006, 139). Because they emerge from a mythological context their speech is non-dialectal. Besides these characters others expose a hybrid embodiment: even though the sovereign Ypsilanti is a Greek warrior for freedom (Hofmannsthal, 2006, 682) his comments show a strong Austrian accent. Also the roman emperor speaks in the Austrian dialect. The differentiation between the vernacular and the Standard German lies in the context of the figures: whether they have a mythological or historical background. Hence the dialectal speech in this context could possibly be connected to the condition of being human. The dictionary for austriacisms thus provides assistance for the analysis of literary texts in two ways. Besides the common function of a reference book to clarify terms, via quantitative evaluation the dictionary can show in which literary genre vernacular terms accumulate and help to analyse the respective motivation for the use of austriacisms. Exemplarily it could be ascertained above that Austrian dialectal forms occur mostly in the dramatic genre. The scenic character with direct speech can thus be evaluated as a criterion for the increased application of the dialect. Other possible indications are historical contexts and social hierarchies.

8 Conclusion

We presented an LMF model for the representation of a lexicon in the humanities computing environment eHumanities Desktop, discussed the peculiarities of dialectal lexica and the Austrian German dialect, described a lexicon creation procedure, which can serve as a role-model for other NLP dialect ressources. This lexical ressource (Digitized Austrian Lexicon Supplement (DALS)) is now available openly. We showed how the resource can be used for philological or linguistic

analyses by example of the historical corpus of Austrian German by Hugo von Hofmannsthal.

8.1 Collaboration

The successful cooperation between the humanities and computer science repeatedly involves such decisions in NLP tasks, where corpus-suited development of methods and error rates must be counterweighted against the effort of manual labour, trying to collaborate in the most cost efficient way possible. With very large data, this might be only achievable by application of software, for very small corpora on the other hand, manual labour may - consider the German saying "mit Kanonen auf Spatzen schiessen" (to shoot at sparrows with canons) - be the quicker and more efficient way. Historical corpora due to their size at the border between these two cases may be especially well-suited and fruitful for a digital humanities cooperation. The subsequent steps of the creation of this lexicon may serve as an example.

9 Acknowledgements

The LMF representation and the lexicon have been developed for the eHumanities Desktop in the lab for text-technology (computer science) at Goethe University Frankfurt in collaboration with the Freies Deutsches Hochstift. We would like to thank the federal state of Hesse's LOEWE program which is the financial source for the Schwerpunkt of "Digital Humanities"[16]. Lastly, we would like to thank the authors of the websites from which we extracted our tokens.

References

Auburger, L. (2011). *Boarische Orthographie*. Pro Business.

Back, O., Benedikt, E., Blüml, Karl, E., Jakob, H., Maria, M., Hermann, P., Heinz-Dieter, and Tatzreiter, H. (2009). *Österreichisches Wörterbuch*. öbv, Wien, 41. edition.

Bashaikin, N. (2005). Dialekt im Cyberspace. Überlegungen zu einigen sozio- und pragmalinguistischen Aspekten. In *Bayerische Dialektologie*, pages 439–450. Universitätsverlag Winter.

Beyerl, B., Hirtner, K., and Jatzek, G. (2009). *Wienerisch*. Reise Know-How Verlag.

Budin, G., Majewski, S., and Mörth, K. (2012). Creating Lexical Resources in TEI P5. *Journal of the Text Encoding Initiative*, 3.

Clyne, M. (1992). *Pluricentric languages. Differing norms in different nations.* Gruyter.

Eroms, H.-W. (2005). Relativsatzmarkierung im Bairischen. In *Bayerische Dialektologie*. Universitätsverlag Winter.

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., and Soria, C. (2006). Lexical Markup Framework. *LREC*.

Gardner-Chloros, P. (2009). *Code-switching*. Cambridge University Press.

Geyer, I. (2005). Belegdarbietung in Grosslandschaftswörterbüchern im Spannungsfeld von Zeit und Raum am Beispiel des Wörterbuchs der bairischen Mundarten in Österreich (WBÖ). In *Bayerische Dialektologie*, pages 195–204. Universitätsverlag Winter.

Gleim, R., Mehler, A., and Ernst, A. (2012). Soa implementation of the ehumanities desktop. In *Proceedings of the Workshop on Service-oriented Architectures (SOAs) for the Humanities: Solutions and Impacts, Digital Humanities 2012, Hamburg, Germany*.

Hofmannsthal, H. (1986). *Der Rosenkavalier. Zum Geleit. Sämtliche Werke XXIII - Operndichtungen I*. Ed. Dirk O. Hoffmann and Willi Schuh.

Hofmannsthal, H. (1998). *Die Frau ohne Schatten, Sämtliche Werke XXV.1 - Operndichtungen 3.1*. Ed. Hans-Albrecht Koch.

[16] http://www.digital-humanities-hessen.de/

Hofmannsthal, H. (2006). *Wiener Pantomime. Sämtliche Werke XXVII - Ballette, Pantomimen, Filmszenarien*. Ed. Gisela Bärbel Schmid and Klaus-Dieter Krabiel.

ISO24613 (2005). Language resource management - Lexical markup framework. *ISO Geneva*.

Jussen, B., Mehler, A., and Ernst, A. (2007). A Corpus Management System for Historical Semantics. *Sprache und Datenverarbeitung. International Journal for Language Data Processing*, 31(1-2):81–89.

Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR*, 163(4):845–848. english in: Soviet Physics Doklady 10 (8) (1966) 707–710.

Markhardt, H. (2005). *Das österreichische Deutsch im Rahmen der EU*. Peter Lang.

Mauser, P. (2005). O, là là: Oalala und Schalala. Diminution im südmittelbairischen Dialekt des Salzburger Lungaus. In *Bayerische Dialektologie*. Universitätsverlag Winter.

Mauser, W. (1982). Österreich und das Österreichische in Hofmannsthals "der Schwierige". *Recherches germaniques*, 12:109–130.

Mehler, A., Gleim, R., Waltinger, U., Ernst, A., Esch, D., and Feith, T. (2009). eHumanities Desktop — eine webbasierte Arbeitsumgebung für die geisteswissenschaftliche Fachinformatik. In *Proceedings of the Symposium "Sprachtechnologie und eHumanities", 26.–27. Februar, Duisburg-Essen University*.

Mehler, A., Schwandt, S., Gleim, R., and Jussen, B. (2011). Der eHumanities Desktop als Werkzeug in der historischen Semantik: Funktionsspektrum und Einsatzszenarien. *Journal for Language Technology and Computational Linguistics (JLCL)*, 26(1):97–117.

Meibauer, J. (2007). *Einführung in die germanistische Linguistik*. J.B. Metzler.

Niebaum, H. and Macha, J. (2006). *Einführung in die Dialektologie des Deutschen*. Max Niemeyer Verlag.

Rowley, A. A. (1990). *The dialects of modern German*, chapter 14 - North Bavarian, pages 417–438. Routledge.

Thomason, S. G. (2001). *Language Contact - An Introduction*. Georgetown University Press.

Waltinger, U. (2010). *On Social Semantics In Information Retrieval: From Knowledge Discovery to Collective Web Intelligence in the Social Semantic Web*. PhD thesis, University of Bielefeld.

Wandruszka, A. (1967). Das Zeit- und Sprachkostüm von Hofmannsthals "Rosenkavalier". *Zeitschrift für deutsche Philologie*, 86:561–570.

Wiesinger, P. (1990). *The dialects of modern German*, chapter 15 - The Central and Southern Bavarian Dialects in Bavaria and Austria, pages 438–519. Routledge.