Published in: Bański, Piotr/Kupietz, Marc/Barbaresi, Adrien/Biber, Hanno/Breiteneder, Evelyn/Clematide, Simon/Witt, Andreas (eds.): Proceedings of the LREC 2018 Workshop "Challenges in the Management of Large Corpora (CMLC-6)" 07 May 2018 – Miyazaki, Japan. - Paris: European language resources association (ELRA), 2018. Pp. 26-32

Example-Based Querying for Linguistic Specialist Corpora

Roman Schneider
Institute for the German Language (IDS)
R5 6-16, 68161 Mannheim, Germany
schneider@ids-mannheim.de

Abstract

The paper describes preliminary studies regarding the usage of Example-Based Querying for specialist corpora. We outline an infrastructure for its application within the linguistic domain. Example-Based Querying deals with retrieval situations where users would like to explore large collections of specialist texts semantically, but are unable to explicitly name the linguistic phenomenon they look for. As a way out, the proposed framework allows them to input prototypical everyday language examples or cases of doubt, which are automatically processed by CRF and linked to appropriate linguistic texts in the corpus.

Keywords: Grammar and Syntax, Infrastructures and Architectures, Information Retrieval, Machine Learning Methods, Specialist Corpora

1. Introduction and Related Work

Specialist corpora do not only serve as data foundation for linguistic studies. Regarding the content aspect, they also provide invaluable access to research results reported in the corpus texts, and thus could be used to promote the transfer of knowledge from specialist corpora to individual learners. This relates in particular to monitor corpora of scientific journals, (virtual) collections of reference books for a certain domain, and archives comprising online information systems from the web.

The content of such information systems is sometimes technically stored as plain text, but more often as a combination of semantically structured XML-hypertexts and text-specific metadata. A prominent example for the linguistic domain is the grammatical information system *grammis*, hosted at the Institute for German Language (IDS) in Mannheim (IDS-Mannheim, 2018) (Schneider and Schwinn, 2014). It brings together terminological, lexicographical, bibliographical, and corpus-based information about German grammar, and combines the description of grammatical structures from a syntactic, semantic, or functional perspective with multimedia content such as graphics, sound, and animation (see figure 1). Thus, features of spoken language, the construction of morphological or syntactical structures, and the effects achieved by the transformation of these structures can be immediately illustrated. Other modules provide authentic datasets and empirical analyses of corpus studies on specific language phenomena, plus a scientific description of selected linguistic terminology with reference to corpus-based examples.

Among other things, such large hypertext collections – usually with a multitude of contributing authors – have to deal with terminological variety: The use of different vocabularies (i.e., terms or concepts that are specific for a certain approach) within documents can cause considerable difficulties for systematic content retrieval (Bubenhofer and Schneider, 2010) (Sharma and Mittal, 2016). This seems especially true for fields where theories or even authors tend to name comparable concepts differently, and where terminology often reflects miscellaneous needs of heterogeneous user groups.

As a consequence, a notorious problem of specialist corpora like *grammis* is the identification of appropriate content that suits the concrete question of the current user – often a search for the needle in a haystack. Apart from traditional retrieval utilities – (semantically enriched) full text search, keyword lists, table of contents etc. – we believe that natural language could play an important role in the exploration process, inasmuch as it allows users to gain accurate access to appropriate pieces of information without the need of learning specialized query languages or without the time-consuming task of filling out complex search forms. Moreover, it could offer a way-out in situations where users, due to terminological uncertainties, are unable to name a certain problem or the phenomenon they look for.

Related work exists for the underlying idea: The Linguist's Search Engine (LSE) tried to offer an "intuitive, linguistically sophisticated but user-friendly way" (Resnik and Elkiss, 2005) by adopting a strategy called "Query By Example" [1] for web searches based on POS tagging. TIGER Corpus Navigator allows users to navigate a corpus, based on example sentences that represent abstract linguistic concepts (Hellmann et al., 2010). Most recently, (Augustinus et al., 2012) and (Augustinus et al., 2016) introduced example-based treebank querying as a way to search within annotated corpus resources. They allow users to enter natural language sentences or phrase segments as a basis to search for similar syntactic constructions.

We expand this methodologically highly attractive idea in several ways: First, we apply it not on annotated treebank corpora, but on a heterogenous structured specialist corpus. Since the included texts provide information about natural language phenomena (object of investigation) with the help of natural language (means of communication), they consequently should be explorable with the same means of natural language. Second, we see example-based querying as an ideal way to open up scientific corpus resources to a broader public. Our focus is not restricted to users who lack experience in specialized corpus query languages, but also

[1] This approach should not be confused with a method of the same name for describing a database query strategy, originally developed by IBM (Zloof, 1977).

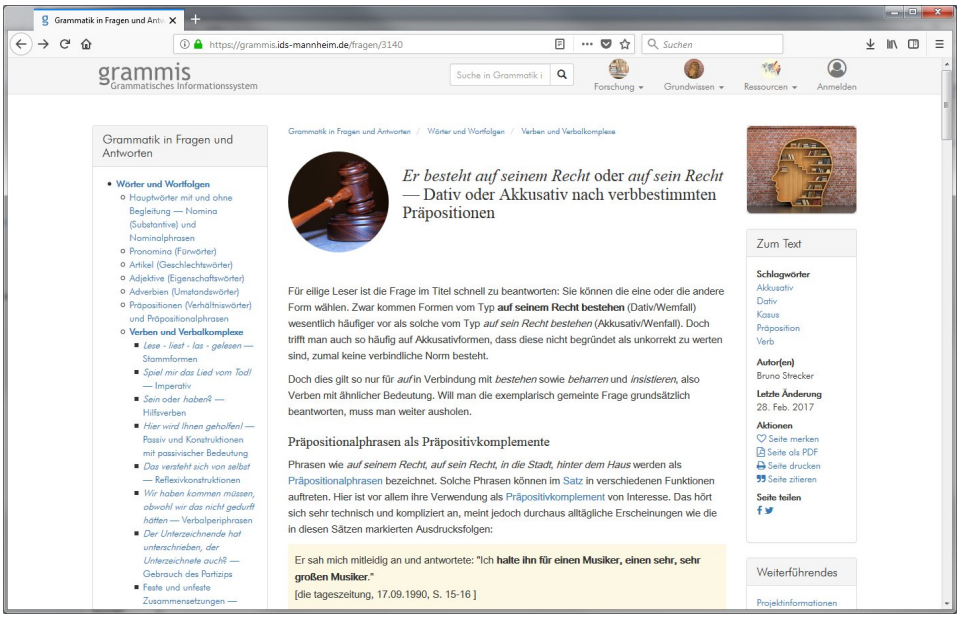

Figure 1: Online access to the grammis specialist corpus

on users with different terminological backgrounds, or even without explicit knowledge of linguistic terminology. The objective is to combine a language-oriented retrieval approach, which is supposed to be suitable for both linguists and linguistic laymen, with a data-oriented foundation.

This framework description is structured as follows: The next section introduces our proposed retrieval layers, as well as the resources they operate on. Section 3. covers the syntax-based layer – the place where example-based querying takes place – in more detail, featuring some prototypical examples. Section 4. summarizes the benefits and gives an outlook on ongoing work.

2. The Retrieval Environment

The *grammis* specialist corpus comprises nearly 3,000 XML-coded hypertext documents on grammatical topics that constitute a 4-million-token collection of specialized language. Furthermore, dictionaries of selected linguistic phenomena (like verb valency, genitive formation, connectors, affixes, prepositions) contribute about 1,500 textual entries with customized XML microstructures and an additional total of 2 million tokens. Both information types are valuable foundations for example-based querying, in the sense that they contain large quantities of natural language examples for illustration purposes, and that they are completely categorized by terminological keywords. Keywords are organized as controlled vocabularies – covering and interconnecting different linguistic theories and schools – within a terminology management system that features ISO-2788/ANSI Z39.19 compliant hyponymy/meronymy relationship types (Suchowolec et al., 2016).

The aggregated 4,500 XML units, containing a mixture of domain-specific specialist language and everyday language example sentences, constitute the overall search space. In order to facilitate content exploration, we consider the following resources:

- **Corpus of Tagged Examples:** Out of the specialist hypertexts and lexical entries, all XML-coded everyday language sentences are added to a corpus database of tagged examples. To enrich these approximately 5,500 samples with POS and morphological annotations about case, number, gender etc., we use the statistical tagger *MarMot* (Mueller et al., 2013), built upon Conditional Random Fields (CRF). For each example sentence, the corpus database stores a back reference to the source document and its corresponding keywords.

- **Dictionaries/Lexica:** The *grammis* lexical resources can be divided into ”flat” dictionaries, organized as simple word lists with attached explanatory texts, and ”enriched” dictionaries with explicitly coded semantic or syntactic content. The latter applies to the *E-VALBU* valency dictionary (IDS-Mannheim, 2010), which is based on the most comprehensive work on German verb valency – *VALBU* (Schumacher et al., 2004) — and includes detailed information about the arguments controlled by a verbal predicate. Furthermore, by adding the onomasiological classification from *Verben in Feldern (VIF)* (Schumacher, 1986), every verb can be assigned to a hierarchical set of semantically founded verb fields and subfields.

- **Terminological Net:** Another semantic resource utilized by the proposed retrieval framework comes out of the *grammis* terminology management system. The approximately 1,400 stored concepts form a poly-hierarchical network of meaningfully related specialized words, using standardised relationship types (syn-

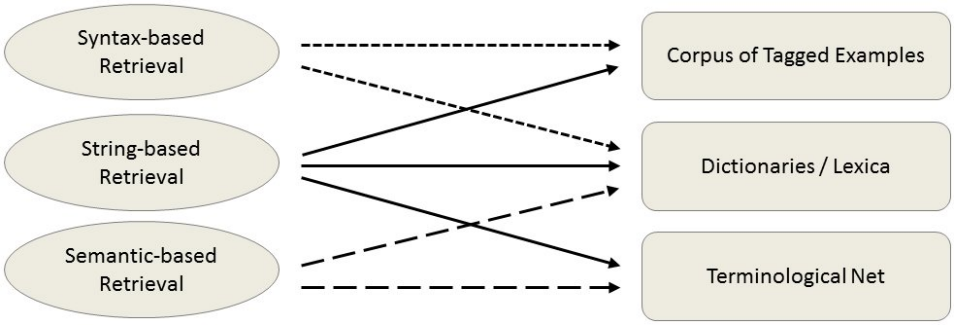

Figure 2: Retrieval Layers and Utilized Resources

onym, broader term, broader term partial etc.). They build the backbone for search strategies that handle user input containing specialist vocabulary. Within this resource, the plurality of linguistic theories is respected, language-specific and cross-linguistic categories are mapped (Haspelmath, 2016) (Moravcsik, 2016).

An ideal platform should allow the users to enter arbitrary free text, typically containing either terminological expressions and/or everyday language examples. This individual input will be automatically processed, using the resources listed above, and linked to corpus content addressing appropriate grammatical phenomena. Search is conducted using three independent layers (string-, semantic-, and syntax-based, see figure 2); their results are rated and merged after completion. [2]

2.1. String-Based Retrieval Layer

The string-based layer delivers reasonable search results in cases where the user enters string sequences that correspond to sequences within the information system's hypertext documents. Syntactically complete sentences as well as shorter phrases or even single words can be processed. The layer operates on the lexical and terminological texts, and looks for exact matches or string-based similarities. The algorithm uses character-based similarity measures, notably (i) Longest Common Substring (LCS), which computes the length of corresponding contiguous characters that exist in search and corpus strings (Gusfield, 1997), (ii) Damerau-Levenshtein for counting the minimum number of operations that are necessary to transform one string into another (Hall and Dowling, 1980), and (iii) Jaro-Winkler, which includes a prefix scale into the computation of number and order of common characters (Winkler, 1990).

[2]Rating and merging of the results from the three layers obviously represents another important procedural issue. We experimented with different weights, and had the impression that the top results of the string- or semantic-based layers in many cases outperformed the results of the syntax-based layer – provided that the former produced reasonable results at all. A best practice evaluation of measures is still pending.

2.2. Semantic-Based Retrieval Layer

This layer potentially adds corpus documents that are semantically related to input keywords, using the document's keyterms – either assigned manually or automatically via Automatic Term Extraction (ATE) (Suchowolec et al., 2017) – and the terminological net (see figure 3). Since terminological concepts are interlinked by multi-hierarchical relations, it is possible to determine semantic similarity by exploiting path length. Important knowledge-based values are the shortest path between two concepts and the maximum depth of the hierarchy (Leacock and Chodorow, 1998). For expanding search within the semantically annotated dictionaries, the layer also takes into account the lexical relations and the hierarchical set of semantically founded verb fields.

2.3. Syntax-Based Retrieval Layer

The syntax layer – which constitutes the heart of the example-based retrieval algorithm – takes over in cases where the user does not formulate his search inquiry terminologically, and where simple word-based lookups yield no satisfactory result. Instead, each user input is regarded as prototypical example sentence or phrase, and undergoes syntax-based processing.

In order to obtain an empirically determined test set, we collect typical everyday language queries, using an anonymized protocol of *grammis'* full text searches. We automatically filter out all requests that contain disambiguated grammatical key terms, and data containing less than three words. Out of the remaining $\sim$ 8,000 sentential expressions, we gradually build up a gold standard, performing manual filtering and double-blind indication of corresponding terminological keywords by human experts. In order to make this collection of typical user queries testable against models trained on the corpus of tagged examples, it is processed with the same morpho-syntactical tagging environment beforehand.

We arrange all computed metadata in a line-oriented CoNLL-like format (Hajič et al., 2009) that can be processed by CRF++ (Kudo, 2005 2013). Some real-life examples from the full text search test set are:

```
Ich PPER nom sg 0 1 0 0 0
habe VAFIN 0 sg 1 pres ind 0 0
mir PRF dat sg 1 0 0 0 0
```

```
oder KON 0 0 0 0 0 0 0
mich PPER acc sg 0 1 0 0 0
in APPR 0 0 0 0 0 0 0
die ART acc sg fem 0 0 0 0
Hand NN acc sg fem 0 0 0 0
geschnitten VVPP 0 0 0 0 0 0 0

sie PPER nom sg fem 3 0 0 0 0
leitet VVFIN 0 sg 3 pres ind 0 0
ein ART acc sg neut 0 0 0 0
pleites ADJA acc sg neut pos 0 0 0
Unternehmen NN acc sg neut 0 0 0 0
```

So after morpho-syntactic annotations are added, the layer operates on the enriched input dataset and tries to identify similar constructions. If this attempt is successful, it can either link directly to the corresponding corpus text, or identify semantically related texts by exploiting the keywords attached to the reference text.
A straightforward approach would initially look for exact syntactical equivalents, and then – if this generates too few results – ignore word order. We believe that the first variant works too restrictively in some situations, and that the second variant is too general and would often produce worthless results. So, if simple decision rules do not help, we argue that recourse to machine learning in general and – since we deal with sequential, word-oriented data potentially containing a broad set of morpho-syntactical metadata – CRF (Lafferty et al., 2001) in particular seems promising. With recourse to statistical methods, it can handle partial matches and situations where a syntax pattern is associated with different targets, and identify appropriate terminological keywords. Other possible metrics for comparing syntactic parse trees would be Tree Edit Distance, Tree Kernels, or Subtree Overlap.

3. Example-Based Querying at Work

We now focus on the example-based retrieval component, evaluating the syntax layer against naturally occurring searches.

3.1. Training

For the computation of a trained model file, we arrange all tagged sentences from the example corpus in the already mentioned line-oriented format. In every line, the first column contains a single token[3], the second column contains the corresponding POS tag, and the following columns represent morphological annotations. The last column shows the ID of an appropriate corpus text[4]:

```
Das ART nom sg neut 0 0 0 0 f3185
Land NN nom sg neut 0 0 0 0 f3185
Niedersachsen NE nom sg neut 0 0 0 0 f3185
wird VAFIN 0 sg 0 ind 3 0 pres f3185
sich PRF acc sg 0 0 3 0 0 f3185
nicht PTKNEG 0 0 0 0 0 0 0 f3185
an APPR 0 0 0 0 0 0 0 f3185
dem ART dat sg masc 0 0 0 0 f3185
europaweit ADJD 0 0 0 0 0 pos 0 f3185
autofreien ADJA dat sg masc 0 0 pos 0 f3185
Tag NN dat sg masc 0 0 0 0 f3185
am APPRART dat sg neut 0 0 0 0 f3185
Freitag, NN dat sg neut 0 0 0 0 f3185
den ART acc sg masc 0 0 0 0 f3185
22. ADJA acc sg masc 0 0 pos 0 f3185
September NN acc sg masc 0 0 0 0 f3185
beteiligen VVFIN 0 sg 0 ind 3 0 past f3185

Sie PPER nom sg fem 0 3 0 0 d312
hoffte, VVFIN 0 sg 0 ind 3 0 past d312
dass KOUS 0 0 0 0 0 0 0 d312
sie PPER nom pl * 0 3 0 0 d312
das PDS acc sg neut 0 0 0 0 d312
bis APPR 0 0 0 0 0 0 0 d312
zum APPRART dat sg masc 0 0 0 0 d312
Abend NN dat sg masc 0 0 0 0 d312
erledigt VVPP 0 0 0 0 0 0 0 d312
haben VAFIN 0 sg 0 ind 3 0 pres d312
wÃ¼rde VAFIN 0 sg 0 subj 3 0 pres d312
```

A template file describes which features should be used for the training run. Each line in the template specifies the involvement of certain metadata by addressing its relative position from the current token, e.g.:

```
# Unigram
U11:%x[-2,1]
U12:%x[-1,1]
U13:%x[0,1]
U14:%x[1,1]
U15:%x[2,1]

U20:%x[-2,2]
U21:%x[-1,2]
U22:%x[0,2]
U23:%x[1,2]
U24:%x[2,2]

U29:%x[-2,3]
U30:%x[-1,3]
U31:%x[0,3]
U32:%x[1,3]
U33:%x[2,3]

U38:%x[-2,4]
U39:%x[-1,4]
U40:%x[0,4]
U41:%x[1,4]
U42:%x[2,4]
```

Since each sentence is interlinked with one or more grammatical keywords and with one hypertext back reference, we distinguish between three training variants: (i) the last column contains a concatenation of all terms (ii) the last column contains only one selected term, as for example the highest/lowest ranking concept within the terminological

[3] Since string-based search is covered within a separate layer, we do not use tokens for the model training, and only consider the subsequent morpho-syntactical features.

[4] As described above, these texts have already been classified by linguistic keywords before.

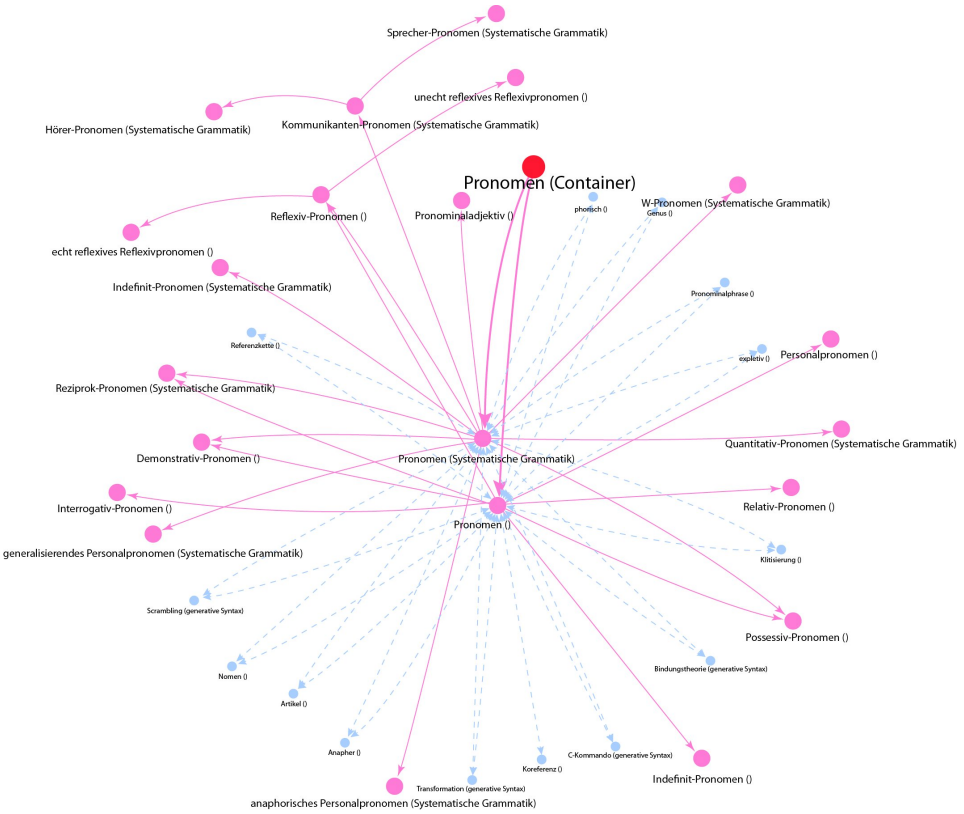

Figure 3: Partial visualization of the terminological net

net (iii) as in the training example above, the last column contains the back reference to a hypertext document, which in turn is annotated with one or more keywords.

3.2. Classification Testing

Out of our test set, we subsequently present two queries and use the trained model for retrieving appropriate explanatory hypertext documents. Work on the best adjustment of the classification model is still in its early stages; we will present a comprehensive evaluation after completion. Nethertheless, we believe that the following examples illustrate the fundamental suitability of example-based querying for the linguistic domain.

3.2.1. Use of Case for Date Specifications

Table 1 shows the tagged input of the first everday language input example (”am Freitag, den 13.”; English equivalent: ”on Friday, the 13th”). Obviously, the underlying – but not explicitely expressed – question concerns the correct use of case within a German date specification: Is the combination of dative and accusative acceptable, or should dative be maintained for the whole phrase (that would then be: ”am Freitag, dem 13.”)?

And indeed, when applying the back reference model as described above, the algorithm references a suitable explanatory corpus text containing similar example sentences. [5] The corresponding keywords of this document are

Token	POS	Case	Num	Gen
am	APPRART	dat	sg	masc
Freitag	NN	dat	sg	masc
den	ART	acc	sg	masc
13.	ADV	acc	0	0

Table 1: First query example as CRF input

Akkusativ (accusative), *Dativ (dative)*, *Datum (date)*, *Deklination (declension)*, *Flexion (inflection)*, *Kasus (case)*.

3.2.2. Use of Genitive Constructions

As a second example (”das Auto von meinem Vater”; English: ”the car of my father”), we choose an authentic user query that a human native speaker would probably classify as somehow related to the use of genitive constructions, although it does not contain any genitives at all (see table 2). A possible genitive construction would be ”meines Vaters Auto”; English: ”my father’s car”.

A syntactically similar example is found within a *grammis* hypertext on the use of the preposition ”von” and dative case, compared to the ”high-order” style of genitive attributes. Consequently, our classification algorithm generates an expedient link to this document. [6] Its classifying keywords are *Attribut (attribute)* and *Genitiv (genitive)*.

[5] https://grammis.ids-mannheim.de/fragen/3185

[6] https://grammis.ids-mannheim.de/fragen/4550

Token	POS	Case	Num	Gen
das	ART	nom	sg	neut
Auto	NN	nom	sg	neut
von	APPR	0	0	0
meinem	PPOSAT	dat	sg	masc
Vater	NN	dat	sg	masc

Table 2: Second query example as CRF input

4. Concluding Remarks

We proposed the intuitive and efficient use of example-based querying for content retrieval on a large collection of specialist hypertexts dedicated to linguistics. Overall, the results of the preliminary studies reveal the attractiveness of this preprocessing step for the thematic exploration of corpora containing natural language example sentences. When combined with string-based and semantic-based retrieval components, the proposed framework can assist users seeking qualified information in situations where they, due to terminological uncertainties, are unable to name the concrete problem they look for. In other words: the approach described in this article helps to transform object-related introductory questions, that are close to everyday language experience, into category-related retrieval questions.

We believe that example-based querying can also play an important role for the search in specialist corpora with different orientations, as long as they contain annotated natural language material whose morpho-syntactical structure is showing some noticeable characteristics. Possible examples range from the CosMov Corpora for Social Movement Research (*www.semtracks.org/cosmov/*) to the Corpus of American Soap Operas (*corpus.byu.edu/soap/*), but comprise even various types of historical corpora. The crucial point here is the corpus extension by metadata: In order to ensure that the syntax-based retrieval layer is able to identify semantically related texts, each corpus text has to be enriched – preferably automatically – with meaningful keyterms. It would be interesting to find out how additional metadata like the results of dependency parsing (Kübler et al., 2009) would enhance retrieval quality. Besides, example-based querying has already been evaluated for human motion data (Kim et al., 2016) and music retrieval using audio and fuzzy-music-sense features (Su et al., 2014).

Depending on keyword complexity and the number of annotation features, the described task also poses theoretical challenges to machine learning researchers, since different classification approaches seem appropriate. If the example sentences are associated with only one (rather general) keyword, queries generate quantitatively more, but mostly far too imprecise results – high recall, but low precision. Associating multiple keywords to every example sentence tends to produce higher error rates. Our tests indicate that using back references to terminologically classified corpus texts can be a satisfying trade-off.

In order to improve the retrieval quality of future *grammis* releases, we are planning to implement the described solution in conjunction with a fundamental extension of the system's content modules. Example-based querying will then be an important (pre-)processing step for the easy-to-use exploration of the large specialist corpus underlying the online information system. The terminological and ML-related resources will be made publicly available in order to foster follow-up research on example-based querying for natural language resources.

5. Bibliographical References

Augustinus, L., Vandeghinste, V., and Van Eynde, F. (2012). Example-based treebank querying. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*. European Language Resources Association (ELRA).

Augustinus, L., Vandeghinste, V., and Vanallemeersch, T. (2016). Poly-GrETEL: Cross-lingual example-based querying of syntactic constructions. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA).

Bubenhofer, N. and Schneider, R. (2010). Using a domain ontology for the semantic-statistical classification of specialist hypertexts. In *Papers from the Annual International Conference on Computational Linguistics (Dialogue)*, pages 622–628.

Gusfield, D. (1997). *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge University Press, New York, NY, USA.

Hajič, J., Ciaramita, M., Johansson, R., Kawahara, D., Martí, M. A., Màrquez, L., Meyers, A., Nivre, J., Padó, S., Štěpánek, J., Straňák, P., Surdeanu, M., Xue, N., and Zhang, Y. (2009). The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, CoNLL '09, pages 1–18, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hall, P. A. V. and Dowling, G. R. (1980). Approximate string matching. *ACM Computing Surveys*, 12(4):381–402.

Haspelmath, M. (2016). The challenge of making language description and comparison mutually beneficial. *Linguistic Typology*, 20(2):299–304.

Hellmann, S., Unbehauen, J., Chiarcos, C., and Ngonga Ngomo, A.-C. (2010). The TIGER corpus navigator. In *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories (TLT-9)*, pages 91–102. Northern European Association for Language Technology (NEALT).

IDS-Mannheim. (2010). *Das elektronische Valenzwörterbuch deutscher Verben*. http://www.ids-mannheim.de/e-valbu/, DOI: 10.14618/evalbu.

IDS-Mannheim. (2018). *Grammis - Grammatisches Informationssystem*. https://grammis.ids-mannheim.de, DOI: 10.14618/grammis.

Kim, D., Jang, M., and Kim, J. (2016). Example-based retrieval system for human motion data. In *2016 6th International Conference on IT Convergence and Security (ICITCS)*, pages 1–2, Sept.

Kübler, S., McDonald, R. T., and Nivre, J. (2009). *Dependency Parsing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Kudo, T. (2005 – 2013). CRF++: Yet Another CRF Tool Kit. Version 0.58.

Lafferty, J. D., McCallum, A., and Pereira, F. C. (2001). Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, pages 282–289.

Leacock, C. and Chodorow, M. (1998). Combining local context and wordnet similarity for word sense identification. In *WordNet: An electronic lexical database*, pages 265–283. MIT Press, Cambridge, MA, USA.

Moravcsik, E. (2016). On linguistic categories. *Linguistic Typology*, 20(2):417–426.

Mueller, T., Schmid, H., and Schütze, H. (2013). Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332, Seattle, Washington, USA. Association for Computational Linguistics.

Resnik, P. and Elkiss, A. (2005). The linguist's search engine: An overview. In *Proceedings of the ACL 2005 Interactive Poster and Demonstration Session*, pages 33–36. Association for Computational Linguistics (ACL). DOI: 10.3115/1225753.1225762.

Schneider, R. and Schwinn, H. (2014). Hypertext, Wissensnetz und Datenbank: Die Web-Informationssysteme grammis und Progr@mm. In Franz Josef Berens et al., editors, *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*, pages 337–346. IDS, Mannheim. http://nbn-resolving.de/urn:nbn:de:bsz:mh39-24719.

Schumacher, H., Kubczak, J., Schmidt, R., and de Ruiter, V. (2004). *VALBU - Valenzwörterbuch deutscher Verben*. Number 31 in Studien zur Deutschen Sprache. Narr, Tübingen.

Schumacher, H. (1986). *Verben in Feldern. Valenzwörterbuch zur Syntax und Semantik deutscher Verben*. Number 1 in Schriften des Instituts für Deutsche Sprache. de Gruyter, Berlin / New York.

Sharma, V. and Mittal, N. (2016). Cross lingual information retrieval (CLIR): Review of tools, challenges and translation approaches. *Information System Design and Intelligent Application*, pages 699–708.

Su, J. H., Wang, C. Y., Chiu, T. W., Ying, J. J. C., and Tseng, V. S. (2014). Semantic content-based music retrieval using audio and fuzzy-music-sense features. In *2014 IEEE International Conference on Granular Computing (GrC)*, pages 259–264, Oct.

Suchowolec, K., Lang, C., and Schneider, R. (2016). Re-designing online terminology resources for german grammar. In Philipp Mayr, et al., editors, *NKOS 2016 Networked Knowledge Organization Systems Workshop. Proceedings of the 15th European Networked Knowledge Organization Systems Workshop (NKOS 2016)*, pages 59–63. http://nbn-resolving.de/urn:nbn:de:bsz:mh39-52982.

Suchowolec, K., Lang, C., Schneider, R., and Schwinn, H. (2017). Shifting complexity from text to data model. adding machine-oriented features to a human-oriented terminology resource. In J. Gracia, et al., editors, *Language, Data, and Knowledge. Springer Lecture Notes in Artificial Intelligence*, pages 203–212. DOI: 10.1007/978-3-319-59888-8.

Winkler, W. E. (1990). String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In Philipp Mayr, et al., editors, *Proceedings of the Section on Survey Research Methods. American Statistical Association*, pages 354–359.

Zloof, M. M. (1977). Query-by-example: A data base language. *IBM Systems Journal*, 16(4):324–343.