Aspects of Linguistic Complexity: A German – Norwegian Approach to the Creation of Resources for Easy-To-Understand Language

1st Melanie Siegel
Hochschule Darmstadt
Darmstadt, Germany
melanie.siegel@h-da.de

2nd Dorothee Beermann
Norwegian Univ. of Science and Technology
Trondheim, Norway
dorothee.beermann@ntnu.no

3rd Lars Hellan
Norwegian Univ. of Science and Technology
Trondheim, Norway
lars.hellan@ntnu.no

***Abstract*—Our project draws on linguistic resources for German and Norwegian in support of initiatives that try to make public language more accessible. We focus on ”Leichte Sprache” for German and ”Klart Språk” for Norwegian. The former refers to usage forms of German which are easily understood also by users of German with a lower competence in text processing, the latter refers to a governance project which has as its aim to aid institutions in their effort to communicate with the public. While different in their goals both initiatives seek to increase the general access to public information. A central concern is to identify the factors which affect language complexity and set up linguistic resources such as parallel text corpora, linguistic rules and terminology databases. Based on these results and resources, we are building software that supports authors and translators, in close collaboration with them.**

***Index Terms*—easy-to-read language, authoring support, linguistic resources, language complexity**

I. Introduction

The free access to information via the Internet does not include population groups who, for example, have difficulty in processing complex texts due to a disability, due to problems in youth that prevented someone from learning to read properly, and due to language learning. ”Leichte Sprache” (LS) is aimed at involving these people in access to information. On the other hand, there are efforts - especially in Norway - to design language in letters and information from authorities in such a way that the texts are understandable not only for trained civil servants or lawyers, but for all citizens, the ”Klart Språk” (KS). In order to compare and combine these two views of a target group-oriented language and then to translate the findings into NLP tools and resources, we have founded a Norwegian-German research group.

II. State of the Art

For LS, organizations for the disabled in Germany such as the ”Lebenshilfe” established language rules for accessible texts ([1]). These rules concern for example the length and complexity of words and sentences. In recent years these rules have been linguistically refined, extended, and implemented in authoring tools such as LanguageTool and Acrolinx ([2], [3], [4]). An empirical study on the effectiveness of these language rules was carried out in the LeiSA study ([5]). Scientists pointed out that there is not only easy and difficult language, but gradations of complexity, whereby the texts are to be written oriented to the target group, and there are more than two target groups ([6], [7]). The Norwegian Klart språk (KS) initiative is an initiative of ’clear writing’ in the Public Sector in Norway - governmental and municipal offices and institutions, and especially in the Law section.[1] Its aim is to avoid unnecessary linguistic complexity in presenting public regulations and laws, and in contrast to LS, KS thus does not have as its principal aim to integrate larger groups of the population into the written medium, but to make communication with the already ’in’-group more efficient. There is thus less focus in KS on rewriting existing texts in a ’simpler’ fashion, and since the subject matters (i.e., the laws, regulations and facts themselves) remain the same, the KS strategies must observe content over style.[2] The German and Norwegian research and development initiatives described here reflect these differences between LS and KS. The sections 3-5 (3 and 4 for German, 5 for Norwegian) will describe the respective initiatives, and in section 6, we highlight areas where the initiatives can join.

The work in this project was partially supported by DAAD, project id 57402159.

978-1-5386-8212-8/19/$31.00 ©2019 IEEE

[1] http://www.sprakradet.no/Klarsprak/om- klarsprak/om-oss/klarsprak-no/, http://www.sprakradet.no/Klarsprak/kommunesektoren/eDifi, http://www.sprakradet.no/Klarsprak/om- klarsprak/om-oss/klart-lovsprak/, and as an example of a guideline, ”Forstått på første forsøk”. Språkprofil, Lotteri- og stiftelsestilsynet. 2012.

[2] The ’content first’ priority even allows for the exploration of alternative ways of communicating content, as when the Skatteetaten supplements a classical definition of the criteria for being a commuter (’Pendler’) with a decision tree kind of application with buttons representing what would be ”if”-clauses in a definitional conditional statement. The reader is in this way lead through steps where each button is introduced by a simple statement, question or command. In this way the linguistic complexity is highly reduced compared with the definitional text.

III. Language Complexity from the viewpoint of LS: Experiments

In the German initiative we first dealt with the question of the nature of complexity of language. For this purpose, we collected texts in LS from the internet that have an equivalent in standard language in order to compare the language structures and vocabulary of the texts. The texts came from diverse sources, such as the bible, news, fairy tales or political programs. We stored the texts in an SQLite database and aligned them document-wise and partly paragraph-wise. We have 351 parallel documents with 30,517 words in the LS and 126,613 words in the standard language documents.

Based on this data, we first created lists of words that appear exclusively in the LS texts or exclusively in the standard texts using the tf-idf algorithm. The result is a list of 1,329 easy words and 1,425 complex words.

A closer look at verbs combines the occurrence of verbs in our database with the frequency in German language in general and the Flesch-Kincaid index. The 10 easiest German verbs are therefore: 'sein', 'haben', 'sagen', 'gehen', 'können', 'machen', 'kommen', 'geben', 'müssen', 'heißen'. These verbs occur in both text sorts, but much more frequently in LS texts. The 10 most difficult German verbs are: 'gestalten', 'leisten', 'ausbauen', 'ermöglichen', 'gelten', 'erhalten', 'verbessern', 'unterstützen', 'fördern', 'stärken'.

The average sentence length in LS texts was 6.5 words as opposed to the average sentence length in standard texts of 13.6 words. For translation, this means that sentences must be shortened by about half. For classification, this means that sentence length is a relevant feature.

In another experiment we investigated whether the readability index of Flesh-Kincaid ([8]) can predict which category the text belongs to. The average value for our LS texts is 56.75, while the average value for the parallel standard texts is 44.75. Thus, our standard texts are theoretically difficult, while LS texts are classified as easier, but still moderate difficult. In order to classify texts - e.g. for a special search engine for simple texts - the Flesch-Kincaid index could be a valuable feature for machine learning of the classification, but the classification cannot be done on the basis of the index alone.

IV. Supporting Authors with Natural Language Processing

Next, we put a focus on the authors / translators and the authoring process of LS. Lists of language rules are available to authors. In addition, there is "Hurraki", a dictionary with explanations of German words in LS. First authoring tools (LanguageTool, Acrolinx) implement LS rule lists so that texts can be checked for language rules. However, tools that have long been standard for language translators are not available to LS authors. There are no terminology databases with expressions in standard language and their equivalents in LS. There are no translation memories either, nor is there any automatic translation.

When using the automatically checkable language rules, we found that it makes sense to be able to select different rules for different contexts and target groups. We have therefore implemented a user interface for an LS check with LanguageTool that allows to select rules.

A closer look at the parallel documents shows that the translation process cannot be automated in the same way as the translation from one language to another. Information is first reduced for translation in LS. On the other hand, information is added, such as explanations for difficult words. We have implemented an automatic text summary, which can be activated prior to the translation process. In collaboration with LS translators, we found that a terminology database for LS would be helpful that could suggest LS variants for difficult words. To create such a terminology database, we used the word lists we had extracted from the parallel documents and OdeNet, a WordNet for the German language[3]. OdeNet - as all WordNets - stores words in sets of synonyms, so-called "synsets". In OdeNet we marked all simple words from our word lists and saved them with their synonyms. Our authoring tool got an extension to mark these synonyms and suggest the corresponding simple words. The resulting dictionary contains 23,000 entries. However, this dictionary had to be revised manually, because some synonyms are strongly context-dependent. For example, every time the word "zu" ("to") appeared in the text, it was suggested to use the alternative "betrunken" (drunk), which is correct in some restricted context, but wrong in most contexts.

V. The Norwegian Klart Språk initiative

Areas of overlap between the LS and the KS initiatives include linguistic features (constructions and types of words) that the KS guidelines advice avoiding and that LS texts more systematically exclude, the array of relevant NLP tools, and digital language resources supporting or being created through the initiatives. The main existing tool for Norwegian is the computational grammar NorSource, an HPSG-based grammar using the LKB platform[4]. Among its areas of application so far is the grammar-checker 'Norwegian Grammar Sparrer' [5], which checks freely chosen sentences of up to 10 words for grammatical mistakes. For some types of mistakes a description of what is wrong and a correct version of the entry are given relative to the mistake diagnosticized. An extension of this tool as an authoring tool also covering certain stylistic and constructional phenomena - grammatical but for some reason less desirable - is conceivable and technically feasible, but care must be taken as to which phenomena to be identified as 'non-advisable'. For instance, in the Norwegian narrative of 'clear language', one warns against the use of passive constructions, abstract deverbal nouns and long composite words. Identifying these in texts automatically is fairly straightforward, but are they always undesirable? For instance, the sentence "Jeg skal ha en kneoperasjon i neste uke" ('I will have a knee operation

[3] https://github.com/hdaSprachtechnologie/odenet

[4] https://typecraft.org/tc2wiki/Norwegian_HPSG_grammar_NorSource

[5] https://typecraft.org/tc2wiki/A_Norwegian_Grammar_Sparrer

next week'), with the deverbal noun "kneoperasjon" ("knee operation"), is quite as natural a possible paraphrase avoiding such a noun, as in "Noen skal operere kneet mitt i neste uke" ('Somebody will operate my knee next week'). The latter in addition would have funny connotations as to who would do it, and avoiding them would require use of the passive form of the verb, another 'non-desideratum', as in "Kneet mitt skal opereres neste uke" ("My knee will be operated next week"). Thus, at the current stage, it should first be thoroughly investigated what construction features seem allowed in text counting as 'clear', and to this effect we are currently extending the Norsource parser to text from Personalhåndboka ('the Personnel Handbook'), an assembly of regulations concerning personnel policy in companies and institutions, texts which are frequently consulted and thus 'user-proofed' on a large basis. Comparative texts - texts which exist first in a 'complicated' version and then in a 'renewed' version, hardly exist, as indicated above, but many of the institutions (like Skatteetaten, the Norwegian Tax administration) offer example texts from before the KS initiative started and from after, from which one will be able to automatically derive further data that allow to extract constructions and lexical items perceived as 'difficult', and to compare them with those that are perceived as 'clear'. A large analyzed corpus of 22,000 sentences drawn from the Leipzig Wortschatz collection[6] is also being used; these are not sorted according to a preconceived KS-non-KS distinction, but are nevertheless indicative of what passes as normal text. In the labeled TypeCraft corpus for sentence construction types and valence,[7] we for instance find that more than 10 percent of all the sentences contain a passive, and on a cursory look, they feel perfectly fine in terms of clarity and complexity.

VI. Common areas

It is thus clear that the initiatives being compared for German and Norwegian have two aspects of asymmetry - one being that the German initiative was started some years before the Norwegian one, with partially different stakeholders involved and partially different aims, and the other being that the subject matter of "Klart språk", being in relevant respects different from that of "Leichte Sprache", also requires partially different methodologies. Nevertheless, as already stated, there are interesting areas of overlap, one concerning the linguistic features (constructions and types of words) addressed, and to some extent the linguistic tools that can be put to use. The largest overlap are most certainly the digital language resources that must be built in support of these applications, one of them being a comparative verb valence overview and database. Thus, both initiatives can select verbal valence frames from the corpora mentioned and make them publicly available, for instance through TypeCraft. We can here compare frequencies both of individual verbs in the respective areas and valence frames employed more generally, and peu a peu develop this resource into a general comparative valence database for the two languages.

VII. Summary

In this paper we have described the German initiative "Leichte Sprache" and the Norwegian initiative "Klart språk". Despite different target groups and different languages, there are similar approaches for tools to support authors who write in these language variants. The project aims to establish multilingual linguistic resources such as parallel texts, word lists and lists of verbs and their valences with information on linguistic complexity.

References

[1] Netzwerk Leichte Sprache. (2013) Regeln für Leichte Sprache. [Online]. Available: "http://www.leichtesprache.org/images/Regeln_Leichte_Sprache.pdf"

[2] M. Siegel and C. Lieske, "Beitrag der Sprachtechnologie zur Barrierefreiheit: Unterstützung für Leichte Sprache," *Zeitschrift für Translationswissenschaft und Fachkommunikation*, pp. 40–78, 2015. [Online]. Available: http://www.trans-kom.eu/bd08nr01/trans-kom_08_01_03_Siegel_Lieske_Barrierefrei.20150717.pdf

[3] A. Nietzio, D. Naber, and C. Bühler, "Towards techniques for easy-to-read web content," *Procedia Computer Science*, vol. 27, pp. 343–349, 2014.

[4] C. Maaß, "Leichte Sprache," *Das Regelbuch. Münster: LIT Verlag*, 2015.

[5] "LeiSa-Studie an der Universität Leipzig." [Online]. Available: http://research.uni-leipzig.de/leisa/de/

[6] J. Suter, S. Ebling, and M. Volk, "Rule-based automatic text simplification for German," *Bochumer Linguistische Arbeitsberichte*, p. 279, 2016.

[7] A. Baumert, *Einfache Sprache – Verständliche Texte schreiben.* Spaß am Lesen Verlag, Münster, 2018.

[8] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom, "Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel," Naval Technical Training Command Millington TN Research Branch, Tech. Rep., 1975.

[6] http://corpora.uni-leipzig.de/en?corpusId=deu_newscrawl_2011

[7] https://typecraft.org/tc2wiki/Norwegian_Valency_Corpus