EMBEDDIA

Cross-Lingual Embeddings for Less-Represented Languages in European News Media

Research and Innovation Action
Call: H2020-ICT-2018-1
Call topic: ICT-29-2018 A multilingual Next generation Internet
Project start: 1 January 2019 Project duration: 36 months

D2.4: Multilingual language generation approach (T2.3)

Executive summary

In this deliverable we describe the high-level approach to natural language generation used within the EMBEDDIA project. The approach forms a basis for both news generation (investigated in WP5) and report generation from online news comments (WP3). In addition, we describe our initial work on adding neural processing components to take advantage of the contextual and cross-lingual word embeddings developed in WP1 to improve the fluency and variety of the generated language.

Partner in charge: UH

Project co-funded by the European Commission within Horizon 2020 Dissemination Level		
PU	Public	PU
PP	Restricted to other programme participants (including the Commission Services)	–
RE	Restricted to a group specified by the Consortium (including the Commission Services)	–
CO	Confidential, only for members of the Consortium (including the Commission Services)	–

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825153

Deliverable Information

Document administrative information	
Project acronym:	**EMBEDDIA**
Project number:	**825153**
Deliverable number:	**D2.4**
Deliverable full title:	**Multilingual language generation approach**
Deliverable short title:	**Multilingual language generation approach**
Document identifier:	**EMBEDDIA-D24-MultilingualLanguageGenerationApproach-T23-submitted**
Lead partner short name:	**UH**
Report version:	**submitted**
Report submission date:	**30/06/2020**
Dissemination level:	**PU**
Nature:	**Report**
Lead author(s):	**Leo Leppänen (UH)**
Co-author(s):	**Miia Rämö (UH), Hannu Toivonen (UH), Matej Martinc (JSI), Senja Pollak (JSI)**
Status:	**__ draft, __ final, _x_ submitted**

The EMBEDDIA Consortium partner responsible for this deliverable has addressed all comments received. Changes to this document are detailed in the change log table below.

Change log

Date	Version number	Author/Editor	Summary of changes made
24/05/2020	v1.0	Leo Leppänen (UH)	First draft.
31/05/2020	v1.1	Leo Leppänen (UH), Miia Rämö (UH)	Ready for internal review.
05/06/2020	v1.2	Shane Sheehan (UEDIN), Matej Martinc (JSI)	Internal review.
12/06/2020	v1.3	Leo Leppänen (UH), Miia Rämö (UH)	Addressed comments from internal review.
15/06/2020	v1.4	Hannu Toivonen (UH)	Final UH checks.
15/06/2020	v1.5	Senja Pollak, Matej Martinc (JSI)	Adding relexicalization and readability content.
16/06/2020	v2.0	Leo Leppänen (UH), Miia Rämö (UH)	Final changes prepared for quality management.
16/06/2020	v2.1	Nada Lavrač (JSI)	Quality control.
28/06/2020	v2.2	Leo Leppänen (UH)	Addressed comments from quality control.
28/06/2020	final	Leo Leppänen (UH)	Final version. Ready for submission.
30/06/2020	submitted	Tina Anžič (JSI)	Report submitted.

Table of Contents

1 Introduction

The main objective of the EMBEDDIA project is to develop methods and tools for effective exploration, generation and exploitation of online content across languages, thereby building the foundations for the multilingual next generation internet, for the benefit of European citizens and industry using less-represented European languages. One facet of this effort is the work on Natural Language Generation (NLG).

Natural language generation, in broad terms, refers to the use of automation to express underlying non-linguistic data in natural language (Reiter & Dale, 2000; Gatt & Krahmer, 2018). The work described here – and continued further in work packages WP3 and WP5 – is focused on a specific subfield of NLG, namely *data-to-text* NLG. That is, the system inputs are assumed to be some type of structured data. In other words, we are explicitly not looking to generate text from linguistic resources such as other texts or sound recordings (as these are already excluded by the definition of NLG) or visual materials, such as images or video, which is the other major subfield of NLG.

Task T2.3, which this deliverable reports on, is focused on producing a multilingual approach to natural language generation. Work Package WP3 employs the proposed approach to produce natural language reports on analyses conducted on online news comments, whereas Work Package WP5 employs the approach to automatically produce news articles, or news "blanks" consisting of some initial text which human journalists can then refine by improving its fluency further and by adding additional analysis and background which automation is unable to provide.

The other focus of the work conducted in Task T2.3 – beyond producing the general NLG approach – is to improve the fluency and variability of the language by incorporating neural methods in a manner that retains the transparency and trustworthiness of the larger system. We provide an example of such a hybrid approach by using contextual word embeddings to lexicalize (select words) and relexicalize (replace words to improve variability) as part of the larger generation process. For both methods, we also describe how cross-lingual word embeddings can be used to take advantage of resources from high-resource languages when conducting the (re)lexicalization processess in low-resource languages.

In this report, we first provide a brief overview of how data-to-text NLG has been previously approached in the research literature (Section 2), including a brief analysis of the upsides and downsides of the various approaches. In Section 3 we provide a brief description of how the approach developed herein is to be used in Work Packages WP5 and WP3, and describe how the needs of these applications – and the need to include neural components for improved fluency and variation – suggest a modular pipeline approach to NLG.

The first main contribution of this report is then provided in Section 4, where we describe our modular pipeline NLG architecture that can be enhanced with neural processing modules to produce a hybrid approach to NLG. The second main contribution is provided in Section 5, where we describe some variants of neural components that can be used to improve the fluency and variability of the generated language using word embeddings developed in Task T1.2.

Following these main contributions, we provide a brief enumeration of the outputs associated with the present work (Section 7). We conclude this report (Section 8) by providing some of our thoughts on the success of the work and describing some potential future work to be conducted both within and outside the EMBEDDIA project.

We note that Section 2, providing the background on natural language generation literature, is also incorporated into Deliverable D5.2 in an abridged form, where it is used to motivate a similar requirement analysis conducted from the more specific viewpoint of automated journalism. Similarly, as Deliverable D5.2 already incorporates a in-depth requirement analysis from the point of view of automated journalism, Section 3, 'Requirements analysis' in this deliverable only briefly outlines the results of that analysis and extends it to the non-journalistic application (i.e. the future work that will be conducted in WP3) of the language generation approach.

2 Natural language generation

The research literature on NLG is broad, describing a variety of generation tasks ranging from captioning images (You, Jin, Wang, Fang, & Luo, 2016) to chat bots (Oh & Rudnicky, 2000) to producing status reports about baby patients in Neonatal Intensive Care Units (Portet et al., 2009) to weather reports (Reiter, Sripada, Hunter, Yu, & Davy, 2005; Coch, 1998; Goldberg, Driedger, & Kittredge, 1994) and automated news texts (Leppänen, Munezero, Granroth-Wilding, & Toivonen, 2017). As a consequence of this wide range of applications, the approaches employed in the systems, and also their architectures, have varied significantly. This variance was already significant even before the rise of methods based on neural networks, which have since further increased the variance between different approaches.

In the past, several NLG architectures were described in the research literature as 'consensus architectures', or architectures the research community supposedly agreed were a 'standard' approach. Perhaps the most famous of these is that of Reiter and Dale (2000). While later surveys indicate that these architectures had not truly reached a consensus status even at the time, and even less so with the increased use of neural architectures in the recent decade (Gatt & Krahmer, 2018), these attempts at describing a 'universal' architecture nevertheless helped establish terminology that is used to refer to the various processing steps that need to be taken within the large NLG process.

It is generally agreed that the NLG process involves three large subtasks: content determination, document and sentence planning, surface realization (Reiter & Dale, 2000). These stages correspond to deciding what information the text should contain, how that information ought to be expressed and finally expressing it in the decided upon manner, respectively.

It is notable that the above three-subtask division is not the only conceptual framework employed in academic works. Other works, such as Reiter (2007), emphasize the need for 'Signal analysis' and 'Data interpretation' that precede the planning of the document especially in the data-to-text context. Others use a six-way split, for example as follows (Gatt & Krahmer, 2018, p. 9):

1. *Content determination:* Deciding which information to include in the text under construction,
2. *Text structuring:* Determining in which order information will be presented in the text,
3. *Sentence aggregation:* Deciding which information to present in individual sentences,
4. *lexicalization:* Finding the right words and phrases to express information,
5. *Referring expression generation:* Selecting the words and phrases to identify domain objects,
6. *Linguistic realization:* Combining all words and phrases into well-formed sentences.

Even before the introduction of neural methods for NLG, the various divisions such as the above were primarily conceptual: while systems tended to prefer to separate their processing into separate components along *some* such division to subtasks, the various approaches were so different that no single division can be called a 'consensus architecture' (Gatt & Krahmer, 2018). The advancement of neural processing methods of the last decade has also seen increased interest in architectures that forego any division into subtasks, instead opting for global, unified, approaches to NLG (Seminally e.g. Wen et al., 2015, presenting the neural encoder-decoder approach to NLG). As such, the above divisions are more a conceptual aid than a procedure to "do these things in this order along these boundaries".

The recent review of the NLG field by Gatt and Krahmer (2018) identified that, in addition to whether NLG is achieved in a modular fashion – with subcomponents each dedicated to some variant of the tasks described above – or in a unified manner – for example with a neural encoder-decoder architecture –, the various systems can be characterized in terms of whether they employ manually programmed rules or approaches based on machine learning.[1] Here, it is important to highlight that these two questions

[1] While Gatt and Krahmer (2018) also identify a third category of 'planning-based approaches,' which we skip here in the interest of keeping this survey of the NLG background suitably concise. We do not believe the planning-based approaches (that to our understanding are rarer than the others) affects our analysis in a meaningful fashion.

of architecture and method are considered orthogonal: rule-based systems can be global and unified, and neural approaches can be modular.

Rule-based approaches use handcrafted rules, often derived from either corpus analysis and expert consultations (Gkatzia, 2016) to achieve the NLG task. As a consequence of their robustness, they provide a relatively high *quality floor* and allow for transparent and explainable processing. Furthermore, they allow for manual correction of any mistakes in the processing. It is likely due to these properties that especially newsrooms seem to prefer rule-based systems (See Sirén-Heikel, Leppänen, Lindén, & Bäck, 2019, where all interviewed newsrooms used template-based NLG, a subcategory of rule-based systems). While commercial NLG providers are notoriously secretive of their systems' internals, the few available public source code repositories (e.g. Yleisradio, 2018), private conversations with stakeholders and the lack of any explicit advertisement of neural methods indicates that rule-based methods are, indeed, dominant outside of academia. At the same time, rule-based systems are costly to produce and require co-operation between domain experts and NLG/NLP experts to establish the system. It is especially difficult and costly to add variation into the generated texts. This is unfortunate, given the observation that the reusability of the commercial rule-based systems seems to also be very low, at least insofar as it is seen by the customers of NLG providers (Linden, 2017).

Academic work on the other type of NLG systems has in the recent half-decade been extremely focused on using neural networks. Initial works (e.g. Wen et al., 2015) in neural NLG took advantage of the encoder-decoder architectures that had been previously shown to function very well in machine translation (Cho et al., 2014). As such, these approaches essentially modeled NLG as a translation task between a 'data language' and some natural language. Since then, academic works have been slowly introducing more and more structure into the generation process. For example, Puduppully, Dong, and Lapata (2019) describe a neural architecture with three-staged processing where the first stage selects what content to include in the text, the second stage plans the order the content is presented in and the final stage realizes the textual content, while still retaining the ability to train the network in an unified manner. A similar two-stage approach is presented by, for example, Moryossef, Goldberg, and Dagan (2019). Going even further, some recent works have split the neural model into a full neural pipeline with promising results (Ferreira, van der Lee, van Miltenburg, & Krahmer, 2019).

Compared to the rule-based NLG approaches, the neural approaches have various upsides and downsides. On the positive side, they seem to have a much higher *quality ceiling*. In other words, especially in complex domains, they can reach very good results and produce highly fluent text. They are also faster to build than the rule-based approaches, and the same model architecture can be often reused in another text domain, albeit with the models retrained.

At the same time, on the negative side, the need for training data can be debilitating in some domains and languages (Gkatzia, 2016). The need for training data also effectively limits the automation to mimicking what humans have been doing, where as, for example in journalism, there is significant industry interest in applying NLG to produce texts that humans have traditionally been unable to produce. Even when the training data is technically available, the expected output text is often not aligned with the input data, and thus cannot be used directly for the development of an NLG system (Belz & Kow, 2010). Furthermore, at least in limited domains, even recent neural end-to-end approaches failed to conclusively outperform rule-based approaches (Dušek, Novikova, & Rieser, 2018).

Empirical evidence also suggests neural NLG – even the recent multi-stage variants (e.g. Puduppully et al., 2019) – suffer from a type of overfitting called '*hallucination*', where the system produces output that is not based on the underlying data (Reiter, 2018; Nie, Yao, Wang, Pan, & Lin, 2019; Dušek, Howcroft, & Rieser, 2019). This alone can be fatal for the applicability of neural methods (at least the present state-of-the-art) to some domains. Finally, neural approaches are inherently opaque to inspection, which has significant consequences for trustworthiness and error correcting. As the systems are opaque, their *quality floors* are unknown, and must often be assumed to be relatively low. This is in stark contrast to the rule-based systems which have lower quality ceilings, but relatively high quality floors. With respect to error correction, neural systems do not allow for targeted system modifications to correct for a specific mistake the system is making. Rather, the system can only be trained further – or completely retrained –

with more data. This, together with the unknown quality floor, means that it is very hard to know whether the general system performance has improved or decreased after some problem is ‘fixed’ by retraining. This last problem is complicated by the observation that the most commonly used automated metrics for estimating the output quality of an NLG system correlate imperfectly with human judges (Reiter & Belz, 2009; Liu et al., 2016; Dušek et al., 2018; Gatt & Krahmer, 2018).

Our interpretation of the current state-of-the-art in NLG is that trainable end-to-end approaches are mainly ready for real-world use in situations where there is ample pre-existing training data of high quality and either the produced texts are very short (i.e. scenarios similar to the E2E Challenge described by Dušek et al. (2018)) or if even major mistakes in individual pieces of output are not problematic, but concurrently high linguistic variation in the output is needed for some reason relating to the application domain of the system.

3 Requirements analysis

As noted above, the NLG approach developed in Task T2.3 is to be used in both Work Packages WP3 and WP5. In WP3, the system is used to produce reports on the commenting behaviour observed on online news, whereas in WP5 the intended use is within a system producing news text.

In the case of news automation (the focus of WP5), Deliverable D5.2 identifies requirements for transparency; accuracy; modifiability and transferability; fluency; data availability; and topicality base on Leppänen et al. (2017). Our analysis of the identified requirements and the state of the art of NLG technology led us to conclude that the needs of automated journalism are best served by an NLG approach that is modular and at least partially rule-based, but also incorporates some neural processing, thus resulting in a hybrid approach. We direct readers to Deliverable D5.2 for details of that analysis. The NLG approach developed in this task is also intended to be used to produce reports from online news comments. As such, we will now provide a brief analysis of the degree to which the requirements imposed by the use of NLG in WP3 are shared with the use of NLG in WP5.

As identified in Deliverable D5.2, the requirements for *transparency* and *accuracy* are related, in that it is insufficient for a system to be accurate if it’s actions are not trusted, and that transparency is an important factor in establishing trust in the system. Without trust, the reports produced in WP3 would not be actionable, thus undermining the usefulness of any developed system. As such, we interpret that the language generation tasks in both WP3 and WP5 share high requirements for both transparency and accuracy.

In Deliverable D5.2, we identified that news automation requires systems that are *modifiable* and *transferable*. For the needs of WP3, these requirements, however, are not as significant: there is a clear application domain where the system is to be applied (reporting on news comments) and it is not anticipated that the system should need to be transferred to a completely new domain. At the same time, it would be beneficial for the system to be able to be extended with new analytical capabilities in the future. Our interpretation is that while WP3 might not have quite as high requirements on these requirements, there is neither a specific requirement for *low* modifibiality or transferability. As such, the requirements imposed by WP3 and WP5 are well-aligned.

In terms of *fluency*, the level required is dependent on how the system is intended to be used. While the exact requirements are hard to quantify, it is clear that a minimal level of fluency is needed for the output to be understandable. In the case of both WP3 and WP5, it is insufficient to be correct if the reader misunderstands what they read. As such, these requirements are well aligned.

Finally, in Deliverable D5.2 we identified a need for *availability of data*, noting that white it’s crucial from a business perspective, it’s less important from an academic perspective. For the intended use in WP3, where the system is to be applied in a known domain and the system input data is produced within the EMBEDDIA project, we believe this requirement is not meaningful. Similarly, in terms of *topicality*, we can simply assume that the data used in WP3 is sufficiently topical; if it was not, this task would not have been undertaken.

As a summary, we note that the requirements imposed on the NLG approach by WP3 (reporting on comments) are highly aligned with those imposed by WP5. Where they differ, the requirements imposed by WP3 are less strict but not at odds with those imposed by WP5. In fact, we can view reporting on the online news comments as producing 'meta news reports' about the comments for internal use. Framed in such a way, it is clear to us that it is desirable to produce a single generation approach, driven by the needs of WP5, that can fulfill the needs of both tasks. As such, we conclude that the needs of both WP3 and WP5 are best served by an NLG approach that is modular and at least partially rule-based, but also incorporates some neural processing, thus resulting in a hybrid approach.

4 The EMBEDDIA language generation approach

Based on the above analysis of requirements imposed by the news domain, we believe that a modular pipeline architecture is – given the present state-of-the-art in natural language generation – the most suitable general architecture for news automation. As such, we have selected as a starting point the architecture we previously used in producing news articles on elections (Leppänen et al., 2017). We have further modularized the architecture and iterated on its design.

This work has taken the shape of a near-complete rewrite of the underlying code base. This includes the separation of an *NLG core* module, which provides interface definitions for the various components, domain and language agnostic code that can be shared between various implementations of the architecture and some simple default implementations of components. For example, the core contains the logic for reading and parsing from textual definitions the templates discussed in Section 4.3. Similarly, Implemented as object oriented Python 3 code, the components and interfaces of the core can be extended through inheritance or replaced wholesale to incorporate completely new approaches to the processing made in the various components.

The EMBEDDIA NLG architecture (shown in Figure 1) consists eight primary stages: message generation, document planning, template selection, lexicalization, aggregation, named entity resolution, morphological realization and surface realization. We describe these modules in more detail below.

The modularity of the architecture allows us to employ both rule-based modules and neural (or otherwise machine learning based) modules in the same architecture. As the rule-based and machine learning approaches have complementing upsides and downsides, this hybrid approach allows us to always pick the option that fits best the requirements for any stage of the pipeline.

4.1 Message generation

During *message generation*, system input data is translated into immutable atomic units of information known as *Facts*.[2] The specific fields of the Fact data structure are dependent on the exact domain, but in general it describes either some event or condition at a specific time span. For example, a fact might correspond to the idea that the number of COVID-19 patients in some specific geographic area at a specific time was some specific value. Other, separate, facts would then describe the change that number represents compared to previous times (i.e. the day before, a week ago, a month ago, etc.), what those changes correspond to in terms of rates of increase in the number of patients, etc. The fundamental properties of these Facts are that they represent the true knowledge of the system (and are as such immutable to prevent accidental changes) and are atomic, in the sense that they are the minimal units of information that can either be included or excluded from the text being generated. The Facts are also associated with an estimate of *statistical interestingness*, which indicates the degree to which the fact is an outlier, i.e. how surprising it is compared to a series or set of related facts.

[2] We ignore here the larger epistemological debate of the nature of truth and knowledge. It is simply assumed here that such a thing as objective truth exists and that the input data corresponds to it.

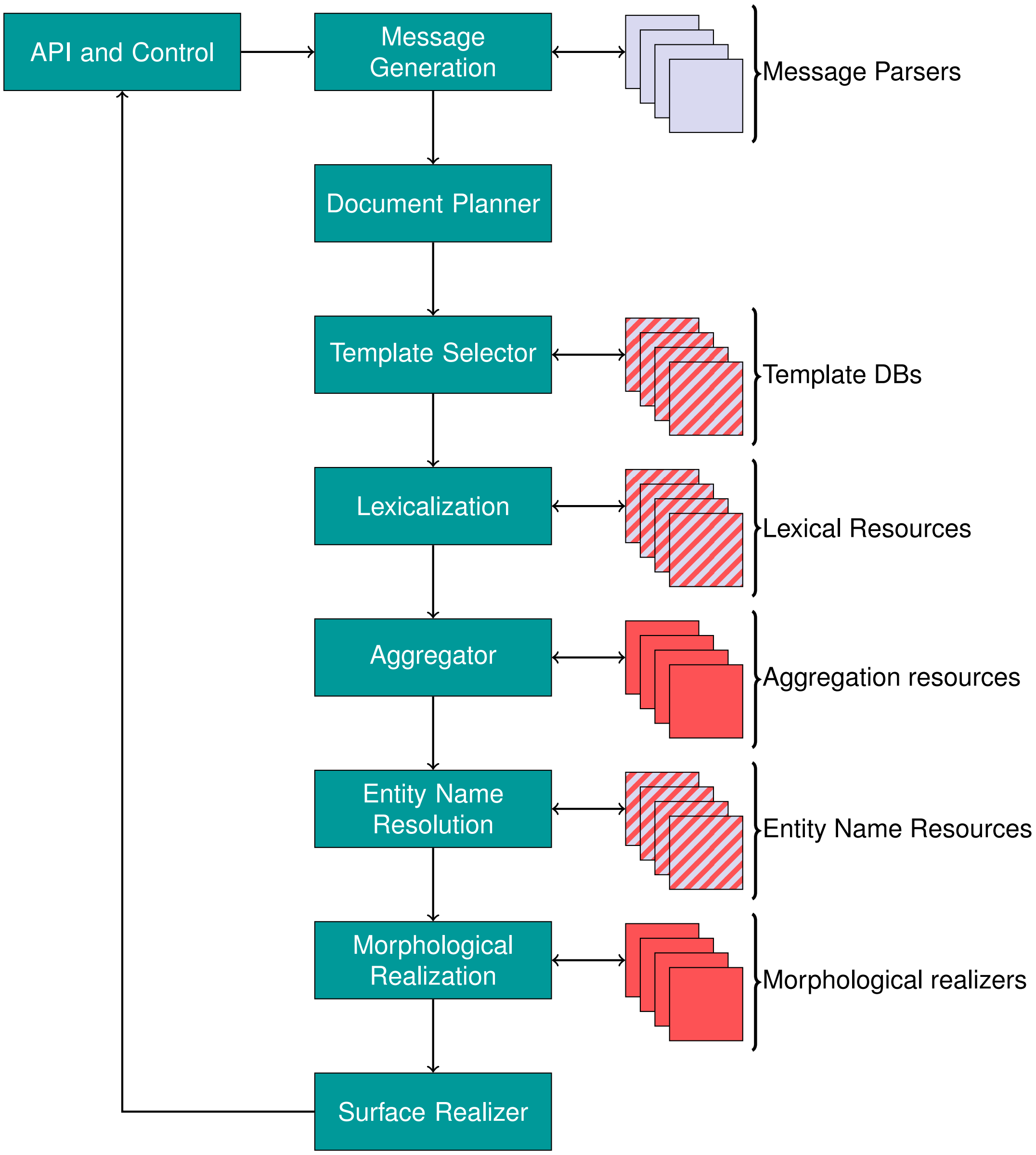


Figure 1: High-level architecture of the EMBEDDIA NLG approach. The overlapping boxes in the right-hand column indicate resources that vary based on either the language or the language and the domain. In this column only, the coloring of the boxes indicates whether the resource in question is dependent on the generation language (e.g. Morphological Realizers), the generation domain (e.g. Message Parsers), or both (hatched boxes).

While the facts themselves are immutable to ensure that they are not accidentally modified during the generation, it is useful to be able to attach mutable information to them. To this end, the Facts are embedded in another data structure, called a *Message*. While there is initially a one-to-one correspondence between the Facts and the Messages, latter stages of the processing can combine multiple facts into a single Message. The Message also contains a field for an estimate of *interestingness* (called 'newsworthiness' in the case of WP5), which contains a numerical estimate of the *value* of the message. This estimate is based on the statistical interestingness of the underlying Fact, but also accounts for the human factors of 'what is interesting'. For example in the case of news automation, when discussing various countries the value might be increased for topics that are of either high societal or personal importance, thus incorporating factors of both news values as defined by the party in charge of the system (i.e. the newsroom) and potential personalization done to account for the reader's preferences. For non-news domains, this factor incorporates the concept of 'how important is it to describe this Fact to the reader.'

In Figure 1, the message generation system is attached to a set of message parsers. These domain-specific but language-independent components provide the logic that allows the system to parse the domain-specific system input and translate it to the Message and Fact data structures. In some cases, such as the EuroStat case study described in D5.2, it is also possible to only have a single, globally employed message parser that then assumes the data is preprosessed into some shared format.

4.2 Document planning

During *document planning*, the system observes all the Messages, which represent what *could* be included in the text being produced, and decides both what information is to be actually included in the text, and in which order it should appear in the text. An important function of this stage is to consider not only the importance of the facts being selected for inclusion, but also their relations to each other: a greedy approach that picks top-*n* most important facts and constructs a document out of them in their order of importance results in an incoherent text that has no meaningful central thread.

As such, the processing conducted at this stage must balance the inherent importance of the various facts as well as the degree to which they form a coherent narrative throughout the text. This means that potentially more important facts can be discarded at this stage for individually less important facts that better fit the other facts of the story. This topic is discussed in great detail in Deliverable D5.3 from the perspective of document planning for news generation. The described method is completely language-independent but assumes that the *type of text* being produced is a news flash style, facts-only report where the temporal element of the described information is not crucial. In other words, the method would be suitable for describing the results of a round of games in a football league, but would not be suited to describing an individual game where the temporal structure of the text is paramount. At the same time, within the types of domains where these kinds of stories are reasonable, the method is largely domain-independent. The modularity of the architecture, however, allows for this approach to be substituted with a domain-specific approach where needed.

The output of the document planner is a *document plan*, a tree-structure that represents the various logical segments of the document. The leaves, at this point, are the individual facts, which are grouped together into paragraph-analogues which together make up the complete document. The rest of the pipeline deals solely with this tree structure, conducting various transformations that step-by-step transfer it closer to natural language text.

4.3 Template selection

The processing thus far has been driven by the input data of the system and has dealt with the content of the document in a language-agnostic manner. Following the structuring of the text in the previous step, the *template selection* stage determines the basic linguistic structures that are to be used to describe the abstract facts in a specific natural language, such as English or Finnish. In our architecture, we assign to each Message a basic template (a skeleton of a phrase) that describes the Fact of the Message.

These templates, provided by humans as a database, are selected so that the Gricean maxim of quantity (Grice, 1975, broadly rephraseable as 'give as much information as is needed, and no more') is fulfilled. That is, when selecting a template, the system must consider the present *contextual assumptions* the reader would have in mind when reading the final text. That is, if two subsequent facts discuss the same location, it is not necessary to mention the location in conjuncture with the latter fact. But similarly, if the next fact discusses a different locale, this change of context must be made explicit to the reader with the use of a phrase such 'in Finland' or the reader will assume the location being discussed has stayed constant. Notably, the templates being chosen at this stage are not complete sentences, but rather a sequence of words interspersed with various *slots* that can be filled with information from the fields of the associated Fact, which in turn are not ready linguistic expressions.

While the templates themselves are, naturally, specific to both some language and some domain, the selection process itself can be abstracted so that a domain and language independent algorithms selects from among domain and language specific templates. This is shown in Figure 1 with the separation of the 'Template Databases' from the 'Template Selector' component.

4.4 Lexicalization

The following stage, *lexicalization* inspects the text for non-linguistic content and replaces it with suitable linguistic expressions. For example, the abstract identifier 'Latest:Confirmed:DailyChange:Abs' might be incorporated into the template by a slot referencing a fact's field. In the context of an English language news report on the COVID-19 situation, this identifier might be expressed as '*the absolute daily change of confirmed COVID-19 cases*' when embedded in a specific context, but could also be expressed as 'the number of new cases' in another. Notably, the lexicalization stage leaves untouched any references to *domain entities*, such as people, time or locations, which are handled separately later as they are often shared – at least to some degree – between domains.

As in the preceding template selection phase, the lexicalization is dependent on resources that are specific to both the language and the domain. For instance, the above COVID-19 related example is tied to both the COVID-19 domain and the English language. Similar resources are needed for all languages and domains. However, as with the template selection process, these language and domain specific resources can be separated from the general algorithms employed and stored separately. Thus, adding support for a new domain or language merely warrants the addition of the new resources and not the modification of the lexicalization component in general.

It is notable that in some scenarios, it might be prudent to conduct the lexicalization either completely or in part after the next stage, aggregation. In such scenarios, the modularity of the pipeline allows for these stages to be swapped, or a second lexicalization stage to be added after aggregation, without affecting the rest of the pipeline.

4.5 Aggregation

Next, the *aggregation* stage inspects the document as a sequence of facts and determines suitable locations where subsequent facts can be more concisely expressed as a single message. Using a simple natural language example, aggregation would observe that the single expression 'The number of COVID-19 cases increased by 19 from yesterday and 98 from this time last week' is more suitable than the same split into two sentences as 'The number of COVID-19 cases increased by 19 from yesterday. The number of COVID-19 cases increased by 98 from this time last week.' We emphasize that the aggregation is not, in reality, done with the fully complete linguistic expressions but using parts of the document plan that are partially abstract and still contain unresolved references to entities in the underlying Facts.

It is also notable that the aggregation process must be conducted carefully to avoid causing misinterpretations. For example, aggregating the sentence pair 'The number of new COVID-19 cases was 15 in Finland. The number of COVID-19 cases was 15 in Sweden.' together into 'The number of new COVID-19 cases was 15 in Finland and Sweden.' is ambiguous with respect to whether the number 15 is the *total* number of cases, or whether both countries separately had 15 new cases, thus bringing the total to 30.

In our approach, aggregation is conducted in a relatively simple manner similar to the examples above. While this simple approach does not need any domain-dependent resources, it does need minimal language specific resources. These minimal resources include information like the fact that the corresponding Finnish word to the English 'and' is 'ja.' This approach was selected as a baseline due to its simplicity, its domain independence and the fact that it's almost language independent. More fluent rule-based aggregation would not only require more – and more complex – language specific resources, but also domain-specific reasoning to detect, for example, whether phrases such as 'despite of', 'while', 'whereas', 'similar to how', etc. are good semantic fits. Using more complex aggregation strategies requires determination is phrase polarities ('something good happened *but* something bad happened'), violations of expected causal or correlative relations ('X happened despite of Y'), etc. Such considerations are extremely domain-specific and would thus be poor fits for the architecture that is intended to be easily transferable between domains. At the same time, the architecture *allows* for arbitrarily complex and domain-specific analysis in situations where the added complexity and cost is deemed necessary or worthwhile.

4.6 Entity resolution

Following aggregation, *entity resolution* determines how the various entities (such as locations or times) should be referred to. The Gricean maxim of quantity (Grice, 1975) means that we should refer to an entity, such as a person, by using a minimal sufficient reference that tells the reader who is being referred to, but does not contain any extranous information. As such, always referring to a person using their full name is a violation of the maxim. Similarly, however, *always* referring to all people using pronouns would violate the maxims as the text would provide insufficient information to deduce who are being discussed. As such, any references to entities must be made considering the previously mentioned entities as well as the context the reader assumes when reading a specific sentence. More concretely, the system should refer to a person using a pronoun in situations where it can be done without danger of the reader misunderstanding the reference. Similarly, dates should be expressed using short expressions such as 'yesterday', 'today', 'this day last week' etc., rather than using complete (alpha)numeric dates.

While this processing is *largely* language-independent, matters such as (grammatical) gender must be considered. For example, in English two different gendered people can be referred to by pronouns in the same sentence without confusion as the pronouns differentiate between genders. At the same time, in the case of Finnish, the use of non-gendered pronouns would make it unclear which pronouns refers to which entity.

To this end, the system needs a set of language-specific resources that describes how to refer to these various entities in specific languages and when various entities are in danger of being confused for each other. These resources are, to some degree, also dependent on the domain, as the domain dictates what resources are needed. For example, a system discussing the various European countries would need to be supplied with the names of said countries in various languages, whereas a system describing a topic local to Germany would presumably need names for various cities and similar areas. At the same time, our experience indicates that these resources are often only weakly related to any *specific* domain, as for example the names of the countries are useful in a myriad of domains. While the level of domain-dependence in this stage is low, we have marked the resources as dependent on both the language and the domain in Figure 1.

4.7 Morphological realization

At this stage, the system has a produced a complete plan of what content to express, in which order and using which words. While this plan can be highly similar to natural language, especially in case of languages such as Finnish the various words that make up the content of the document still need to be realized to their correct morphological forms. This is conducted in the *morphological realization* stage, where language-specific actions are undertaken to inflect words to their correct forms.

While morphological realization is of potentially very significant complexity, especially in the case of languages such as Finnish, it is also a well-studied problem in the natural language processing literature. As this stage of the process is domain-independent, it is possible for the system to employ at this stage the various available 3rd party morphological generators.

While such 3rd party tools are not strictly necessary for languages of relatively low morphological complexity, such as English, they are increasingly important for languages with higher morphological complexity. For example, Finnish nouns have over 2000 distinct morphological forms (Karlsson, 1996). While the majority of them are unlikely to be used outside of very niche situations, even the simpler inflections usually require relatively complicated analysis of the stem to determine which vowels to use.

4.8 Surface realization

Following this, the document is completely in natural language. The only remaining task, *surface realization*, finally translates the tree structured document plan into text that can be provided to the end user. This involves flattening the structure, capitalizing sentences and adding sentence-final punctuation, as well as potentially wrapping the various flattened parts of the tree in markup language. This stage concludes the pipeline, and the resulting output is a flat string of characters that can be provided to the user in whatever format is most suitable. Alternatively, the output can be, for example, a JSON structure that allows for the text to be embedded in a website.

4.9 Notable properties

The notable aspects of the architecture lie in its *separation of concerns*, *modularity*, and *reusability*. We already described separation of concerns and modularity above: as the various stages of the pipeline are clearly delineated and self-contained, they can be modified or replaced as needed. Furthermore, the pipeline allows for new stages to be injected into the middle of the pipeline where needed to conduct various additional tasks without affecting the rest of the pipeline. Below, we exemplify this property by describing one way in which hybrid methods taking advantage of contextual and cross-lingual word embeddings can be used to add varied and more fluent language to the generated texts.

In terms of reusability, the pipeline approach's distinct upside is in the fact that whereas the initial components of the pipeline are domain-specific, the latter parts are decreasingly less so, rather being more specific to the language being generated. For example, morphological realization can be achieved using standard natural language processing tools in many languages. While some of the components are dependent on both the domain and the language – most significantly the template selection and lexicalization – we have abstracted these processes so that the resources specific to the language and the domain are provided separately from the underlying algorithms that apply said resources. As a result, a system implementation that produces textual content about topic T_1 in language L_1 can be easily modified to discuss topics T_2 and T_3 in the same language, as these only warrant changes to the start of the pipeline. Similarly, while providing support for a new language L_2 is more effort than a new domain, once L_2 is supported in one system (or topic), said components can be reused in other systems (or topics).

The above description of the architecture leaves most of the details of any implementation vague, rather focusing on describing the general principles. This is because the same architecture can be applied in several slightly different fashions depending on the precise domain in which text is to be generated. Two case study systems, with slightly different priorities with respect to what *kinds* of extensions are easiest to produce, are described in Deliverable D5.2, which concerns the applications of the above architecture for news generation.

At the same time, the architecture is not limited to producing news, but is applicable to other news-adjacent text genres as well. For example, it can be used to produce natural language reports of analytics data. For example, we intend to employ an implementation of this architecture in WP3, for the purpose of generating reports about online news comments.

5 (Re)lexicalization with contextual and cross-lingual word embeddings

As noted in Section 3, the contexts in which the language generation approach is to be used seem to be best served by a modular, hybrid, pipeline approach. In this section we provide some examples of how neural processing can be included in the pipeline to improve the fluency and variability of the output for both high-resource and low-resource languages.

We first describe a method for improving *lexicalization*, where we add new content words to the phrases produced in an otherwise rule-based manner. Following that, we describe a variation where content words of phrases generated in a rule-based manner are relexicalized, or replaced with other words. For both the lexicalization and the relexicalization method, we also provide modified algorithms that use cross-lingual word embeddings to take advantage of resources available for high-resource languages only even when processing low-resource languages. Finally, we describe results of human evaluations conducted on all the approaches.

In both approaches, we employ directly – without any fine-tuning – the contextual word embeddings developed in Task T1.1 (see Deliverable D1.2). For the low-resource variants, we also employ the cross-lingual word embeddings developed in Task 1.1, again without fine-tuning. While fine-tuning on some highly domain-specific corpus might improve the performance, the results obtained in that manner would not be achievable in situations where such corpora are not available. As such, we believe evaluating the methods on the untuned embeddings provides a better understanding of the realistic worst-case performance that would be observed in most cases.

The content of this section is based on the MSc thesis work of Miia Rämö.

5.1 Lexicalization for improved fluency

Adding content words to phrases generated in an otherwise rule-based manner allows, for example, the addition of modifiers such as adjectives. These modifiers add additional flavour, fluency and variation to the produced texts. This is needed as especially the lack of variation is often seen as a problem in rule-based NLG.

The contextual word embeddings developed in Task T1.1 provide an excellent method for intelligently selecting what words to incorporate into the text. Standard word embeddings represent each token as a static vector. As the vectors are specific to the tokens (i.e. spelling), the vectors are shared between homogprahs and polysemous words, such as the words bank in the sense of ‘a financial institute’ and ‘the side of a river.’ In contextual word embeddings (most prominently Peters et al., 2018; Devlin, Chang, Lee, & Toutanova, 2019) a token’s vector representation is based on the surrounding tokens, its context, and thus encodes its semantics on a much more fine-grained level. In effect, contextual embeddings use the contexts of the different word meanings to distinguish between, for example, the meanings of word ’bank’ in the sentences “I went to the bank to withdraw cash” and “I fell from the bank while fishing.” For additional detail, see deliverable D1.3.

Usefully for our purposes, BERT (Devlin et al., 2019) is pre-trained as a masked language model. During pre-training, the model is given sentences where some words have been masked, and the goal of the model is to predict what the masked tokens were before masking. As such, the 3rd party library used to train the contextual word embeddings in WP1 directly provides facilities to employ them as masked language models to query the model for the most likely tokens to replace a masked token in a specific context. This allows us to generate new content words to target sentences by simply adding mask tokens to the places where the new words are to be added and then querying the masked language model for the most likely words. This naive approach, however, provides for no real control over the generated words and is thus not a reasonable approach in domains such as news where accuracy is paramount.

The most trivial method for making this approach controllable would be to employ a list of allowed words. During generation, the likelihoods of the candidate words would then be observed using the language model and the most likely word selected. A slight modification of sampling from the list of allowed words based on the likelihoods would produce more varied results with some reduction in fluency. This method, however, suffers from a need to predefine these allowed words manually beforehand. It would be preferable to conduct some more intelligent pruning of the tokens suggested by the language model.

This pruning might take the form of, for example, ensuring that the generated tokens are of a handpicked part of speech. In case of many languages with ample linguistic resources, such a check is trivial, as one can simply generate the candidate tokens using the masked language model and then filter based on some 3rd party part of speech tagger. If no suitable replacements are found, the slot can be realized as an empty string. This approach is shown in Algorithm 1. We emphasize that this is not the only possible approach, but simply demonstrates how other linguistic resources can be integrated into the process.

Algorithm 1 Pseudocode describing the method for adding new words to sentences generated using rule-based methods. The approach is tailored for high-resource languages, such as English, and uses additional linguistic resources (here, a part of speech tagger) to conduct further filtering.

```
function LexicalizeWithPosFilter($Sentence$, $PosTag$)
    $ProposedWords \leftarrow$ Top $k$ words proposed by BERT for masked word in Sentence
    $TaggedWords \leftarrow \{(w, \textsc{PosTag}(w)) | w \in ProposedWords\}$
    $FilteredWords \leftarrow \{w | (w, tag) \in TaggedWords \land tag = PosTag\}$
    return Sample($FilteredWords$)
end function
```

The modularity of the language generation approach described above enables us to implement the procedure as follows: Templates, residing in the Template Databases (‘Template DBs’ in Figure 1) are enhanced with placeholder tokens that indicate places where new words are to be generated. A new module is appended after Template Selection, which then runs Algorithm 1, generating the new tokens. Note how the changes are limited to the Template Database and the introduction of a new module in the pipeline. As such, no other modules need to be changed to accommodate the addition of this new lexicalization step.

The approach shown in Algorithm 1 depends on the ability to conduct part of speech tagging. While it is reasonable to assume that 3rd party libraries for this purpose are available for the larger languages, our approach should also work for low-resource languages where suitably high-quality linguistic resources are not available.

In such cases, the procedure `PosTag(·)` can be modified to use the cross-lingual word embeddings to take advantage of the high-resource language’s resources. In the simplest case, given some word W^{Low} in a low-resource language L^{Low}, the cross-lingual word embeddings of language L^{Low} and a high-resource language L^{High} can be queried for the closest word W^{High} in L^{High}. This word W^{High} can then be processed using the linguistic resources available for the high-resource language L^{High} and the results applied to the original low-resource language word W^{Low}. While this process is noisy, it provides at least *some* access to linguistic resources not otherwise available in the low-resource language L^{Low}.

5.2 Relexicalization for more variation

The above method for lexicalization is suitable for generating completely new words, but does not suffice for cases where the existence of *some* suitable word is necessary for the text to be meaningful. For example, it would be suitable for generation of the word *significantly* in the context “Unemployment rose *singificantly* in June 2020”, as the sentence would be meaningful even without the word present. At the same time, it would be unsuitable for generating the word “rose”, as leaving the word unrealized does not produce correct language.

Furthermore, the method suffers significantly from *antonymity*: in the first case, both the words *significantly* and *marginally* would be believable generations for the context, but presumably both would not fit the same underlying data. The same applies with more significant results to words *rose* and *decreased* in the same context.

The first problem – of failing generation – can be solved simply by including in the mask a *seed word* that provides a fallback option to use in case the lexicalization fails. This turns the process from lexicalization in to *re*lexicalization, as rather than generating new words, we are attempting to replace an existing word with an alternative. This seed word, provided manually by the humans, can also be used to deduce which of various candidate words are suitable for the context: if the seed word is the verb “rose,” then it follows that the word “increased” – a synonym – is a suitable replacement, whereas the verb “decreased” – an antonym – would not be.

In the case of high-resource languages, existing language resources on synonymity can be used to deduce which of the candidate words provided by the language model are antonymous and which are synonymous. For example in the case of English, WordNet (Miller, 1995) can be queried for synonyms of the seed word and the candidates retrieved from the synset can then be scored for contextual suitability using a contextual language model. This approach is shown as pseudocode in Algorithm 2.

Algorithm 2 Pseudocode describing a method for relexicalizing using a combination of a language model (based on contextual word embeddings) and a synonym set, such as provided by WordNet.

```
function ChooseReplacementUsingSynset($SeedWord, Context$)
    $Candidates \leftarrow$ GetSynonyms($SeedWord$)
    $CandidatesAndScores \leftarrow \{(c, P_{LM}(c, Context)) | c \in Candidates\}$
    $ReplacementWord \leftarrow$ Sample($CandidatesAndScores$)
    return $ReplacementWord$
end function
```

In the case of low-resource languages, it should be possible to employ a similar roundtrip to that described above with the lexicalization algorithm. Using the cross-lingual embeddings to retrieve the closest word W^{High} in the high-resource language L^{High}, it should then be possible to retrieve potential synonyms S_i^{High} of W^{High} that could then be ‘translated’ back to the low resource language L^{Low} using the same cross-lingual word embeddings to retrieve the analogues S_i^{Low} in the low-resource language L^{Low}. It is however notable that the quality of this method is highly dependent on the quality of the cross-lingual word embeddings used, and that in this first iteration we naïvely left the task of ruling out words with wrong semantics provided by WordNet to the contextual language model.

Like above, this process is relatively noisy and thus does not result in perfect output. Our trials with Finnish (used to simulate a Low-resource language) indicate that at least in the case of languages with high morphological complexity, tools for morphological analysis can be used to significantly improve the results. Using morphological analysis tools, the original morphological form of the word W^{Low} can be stored and the word then lemmatized. Conducting the roundtrip to the high resource language in this lemmatized form then results in lemmatized low-resource language synonyms S_i^{Low}, which can be processed into the morphological form of the original seed word using the same analyzer.[3]

Intersections of embeddings

The relexicalization procedure could further benefit from the method proposed by JSI (Vintar, Simeunović, Martinc, Pollak, & Stepišnik, 2020, work only partly done in the scope of EMBEDDIA project; see Appendix A). The study deals with automatic extraction of words expressing a specific semantic relation by using intersections of word embeddings (note that currently we use static embeddings, but the work could be adapted to contextual embeddings).

The initial assumption in the study was that the word embeddings of a set of adjectives expressing a specific semantic relation share a certain semantic component which can be used to extract other adjectives expressing the same relation. To test this assumption, FastText embeddings (Bojanowski, Grave, Joulin, & Mikolov, 2017) were trained on the multilingual TermFrame corpus (Vintar et al., 2020) made of articles about Karstology, i.e. the study of karst (a type of landscape developing on soluble rocks such as limestone, marble or gypsum) with adjectives expressing selected relations, such as CAUSE, FORM or COMPOSITION.

For each seed adjective expressing a specific semantic relation, the embeddings were used to find a set of 100 closest words according to cosine distance. In order to find words of similar semantic provenance that express a specific semantic relation, in the next step all non-empty intersections between these sets of 100 closest words were calculated for all possible subsets of a set of adjectives for each relation. These subsets ranged in size from 10 to 2, since 10 is the largest subset of seed adjectives for a relation, for which a non-empty intersection was returned. All words found in these intersections are retained as candidate words that express a specific relation and are used in manual evaluation, which showed showed a positive linear correlation between the subset size and precision of the method in the majority of cases.

The approach for acquiring new words expressing a specific semantic relation could be easily transferred to the problem of relexicalization described above, i.e., replacing an existing word with an alternative. By using template filler words provided manually by the humans as seed words, we could generate new words by intersecting the sets of nearest neighbours of these seed words. The hypothesis is that a set of seed words related by a semantic relation of synonymy, share a certain semantic component, that could be leveraged for generation of new synonyms and therefore solve the antonymity problem described above.

5.3 Evaluation

The approaches described above were implemented in a real NLG system producing reports from EuroStat data (See Deliverable D5.2 for a description of a later version of this system). The system produced content in both English and Finnish, allowing us to use the high-resource variants with the English language, and simulate the low-resource variants with Finnish.[4] In all cases, the `Sample(·)` chose the final selection randomly from amongst all candidate words whose likelihoods according to the contextual language model were over a minimal threshold value (0.0005) chosen experimentally.

For the lexicalization method, online judges (both Finnish and English natives) were shown pairs of sentence variants. In each variant, sentence 1 was produced by a rule-based approach and sentence 2 was sentence 1 but with a word added using the method described in Section 5.1. For English, we randomly selected sentence pairs from a set of 50 so that each sentence pair was evaluated by

[3] Many morphological analyzers are constructed as finite automata and can thus be reversed to function as morphological generators.

[4] Despite its low number of speakers, Finnish has a surprisingly high amount of linguistic resources and may better be described as a ‘medium resource language’.

three judges. For Finnish, due to limitations in the platform where the evaluation was conducted, all 21 judges were shown the same 10 sentence pairs. The same procedure was repeated for the relexilization method, with sentence 2 formed by relexicalization rather than lexicalization.

The judges were then asked to rate the following statements on a 7-step Likert scale, with answer 1 signifying 'strongly disagree' and answer 7 signifying 'strongly agree'. The value 4 was a neutral option, labeled 'Neither agree nor disagree':

Q1: Sentence 1 is a good quality sentence in the target language.

Q2: Sentence 2 is a good quality sentence in the target language.

Q3: Sentences 1 and 2 have essentially the same meaning.

For each sentence pair, the judges were then shown two groups of words, 1 and 2. Group 1 contained the word that was added to form sentence 2 from sentence 1. It also included other potential candidates that were not selected during the generation. Group 2 contained words that were ruled out by the system in the filtering step. The following questions were then asked:

Q4: How many of the words in word group 1 could be used in the marked place in sentence 1 so that the meaning remains essentially the same?

Q5: How many of the words in word group 2 could be used in the marked place in sentence 1 so that the meaning remains essentially the same?

The judges answered questions 4 and 5 using a 5-step Liker scale using the following scale:

1. None of the words
2. Less than half of the words
3. Half of the words
4. More than half of the words
5. All of the words

5.4 Results

Results for the high-resource language lexicalization approach, shown in Figures 2 and 3, indicate that the modified sentences were seen as being of at least equal, if not better, quality compared to the unmodified sentences. At the same time, it seems that the respondents believed the sentence meanings remained the same. At the same time, the responses to Q4 and Q5 (Figure 3) indicate that many potentially rather suitable words were excluded by the filtering and similarly many potentially unfitting words were unnoticed by the filtering. This, in turn hints that better results could be obtained by modifying the method by which candidates are selected, but also that it is possible that the results in Q1-Q3 were simply 'lucky' in terms of the specific sampling method used.

For the low-resource variant (which conducts a round-trip to a high-resource language using the crosslingual word embeddings as decribed above), the results shown in Figures 5 and 6, indicate that the modifications did partially compromise the quality of the sentences. Post-hoc analysis indicates that this is likely a problem with the complex morphology of Finnish, which was used as a stand-in for a low-resource language. It is also notable that, according to the judges, the sentence meaning changed significantly in some cases. On the other hand, Figure 6 indicates that most of the 'approved' words were not suitable for the context. This in turn indicates that by improving the word selection, the results might be significantly improved.

At the same time, the results of the *best* POS-tag as figured in post hoc analysis (adverb, see Figure 5b) were significantly better than the aggregate results. Indeed, the aggregate results seem to have been skewed significantly by bad performance in the case of some POS tags. Future work needs to

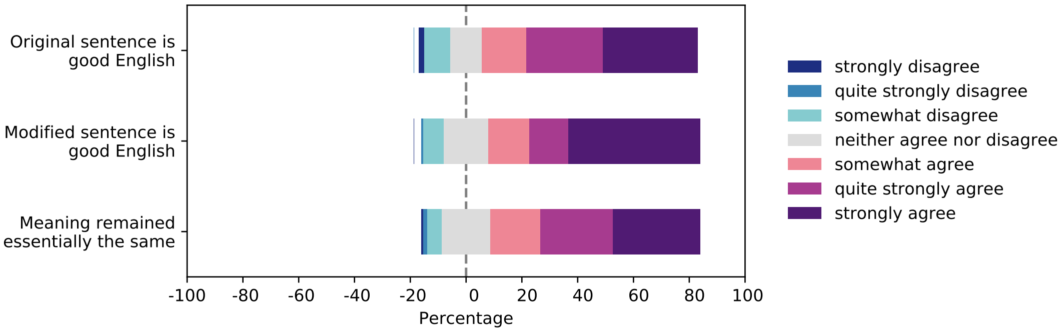


Figure 2: Results for questions 1–3 from applying the high-resource lexicalization approach to English.

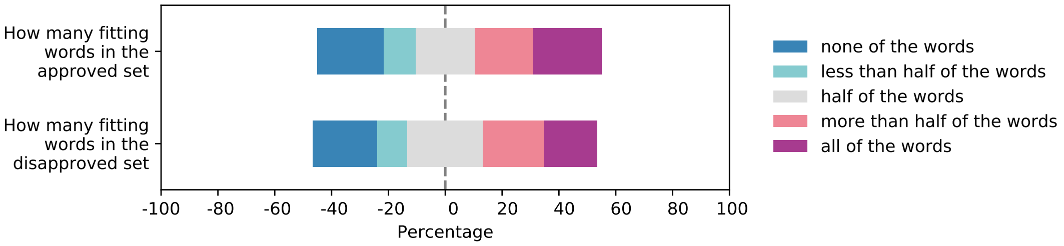


Figure 3: Results for questions 4 and 5 from applying the high-resource lexicalization approach to English.

be undertaken to better understand whether this is a one-off result or indicative of a larger trend with regard to the method's suitability for generating certain POS tags. These results also demonstrate that the round-trip through the high-resource language needs not add prohibitive amounts of error.

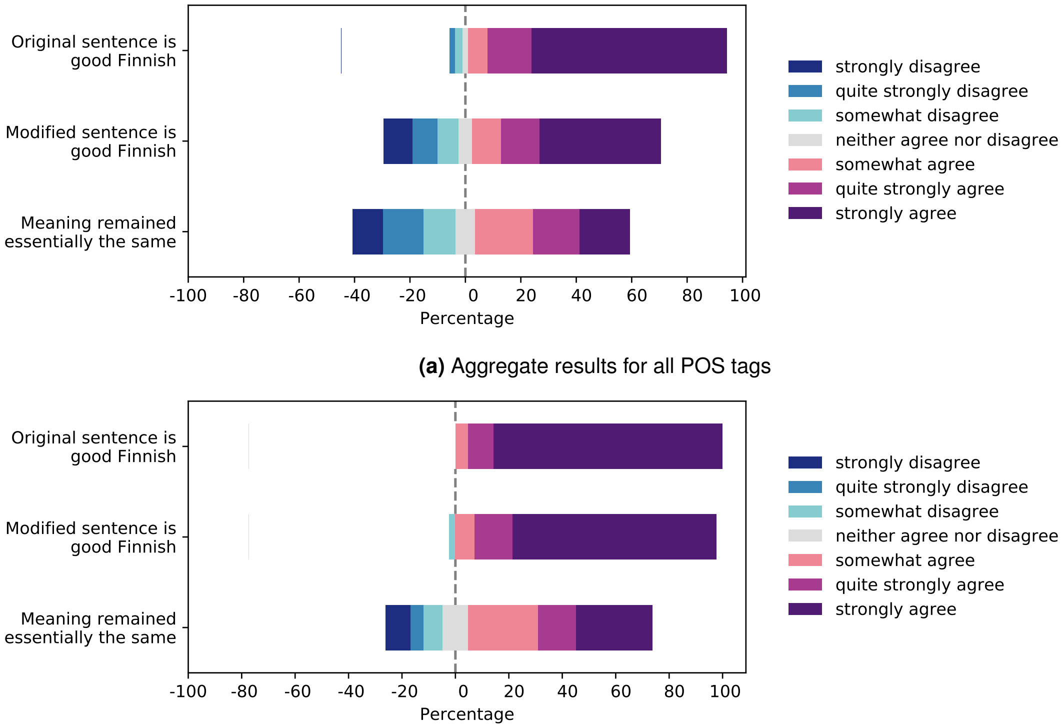
Examples of the lexicalization method's outputs are shown in Figure 4. We note that even the sentence (c), which was scored low, lexicalized with a semantically suitable token and the error stems from its grammatical suitability. The sentence would be made grammatical and meaningful by inflecting '*omaa*' as '*omasta*'.

a) In May 2018, however the growth rate on previous month was for the category housing, water, electricity, and gas and other fuels 0.6.

b) Toisaalta Ruotsissa vuonna 2018 55-64-vuotiaiden naisten tulojen mediaani oli 313792 paikallisessa valuutassa ilmaistuna.

c) Maltalla vuonna 2015 kotitaloudet maksoivat *omaa terveydenhuollostaan itse 37.47 %.

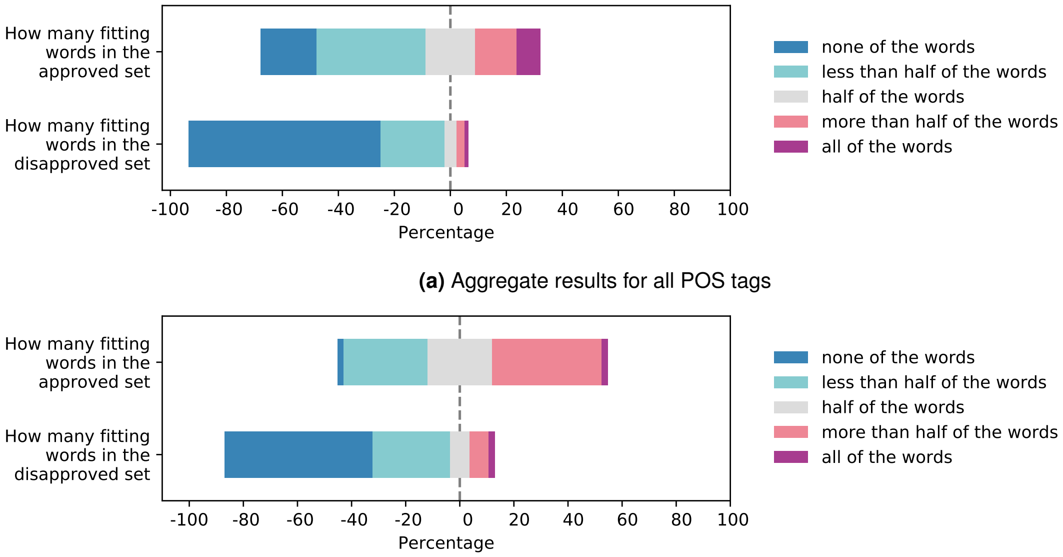
Figure 4: Example sentences produced using the lexicalization method in English (a) and Finnish (b, c). The underlined tokens were added during lexicalization. Sentences (a) and (b) were scored high by the judges. Sentence (b) translates as '*On the other hand, in Sweden in 2018 the median income of females between ages 55 and 64 was 313792 when expressed in national currency.*'. Sentence (c) was scored low and contains an ungrammatical token denoted by *. It translates roughly as '*In Malta in 2015, households paid out-of-pocket 37.47 % of their *selves own healthcare.*'



(a) Aggregate results for all POS tags

(b) Results for best POS tag, RB, or adverb

Figure 5: Results for questions 1–3 from applying the low-resource lexicalization approach to Finnish, with English as the high-resource language. Figure (a) presents the aggregate results over the various parts of speech which were trialed, where Figure (b) presents the results for what was identified as the best POS in post hoc analysis.



(a) Aggregate results for all POS tags

(b) Results for best POS tag, RB

Figure 6: Results for questions 4 and 5 from applying the low-resource lexicalization approach to Finnish, with English as the high-resource language. Figure (a) presents the aggregate results over the various parts of speech which were trialed, where Figure (b) presents the results for what was identified as the best POS in post hoc analysis.

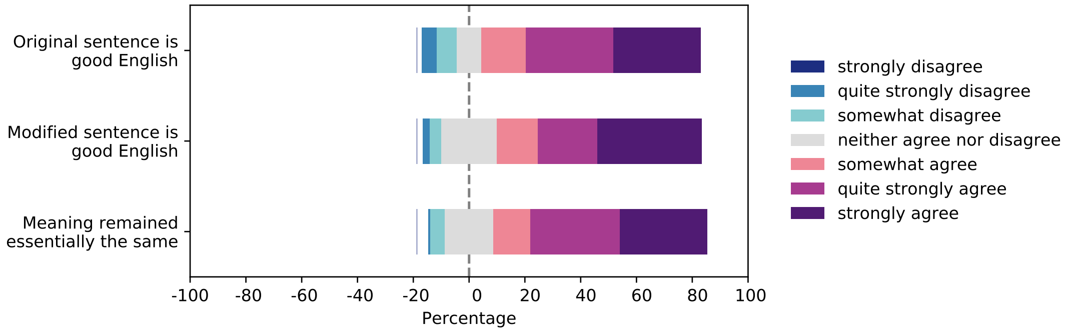


Figure 7: Results for questions 1–3 from applying the high-resource relexicalization approach to English.

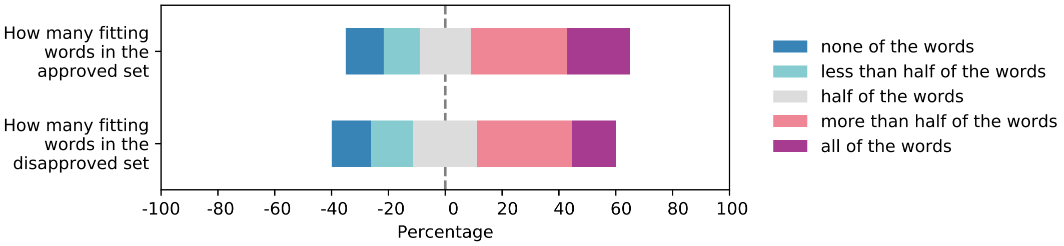
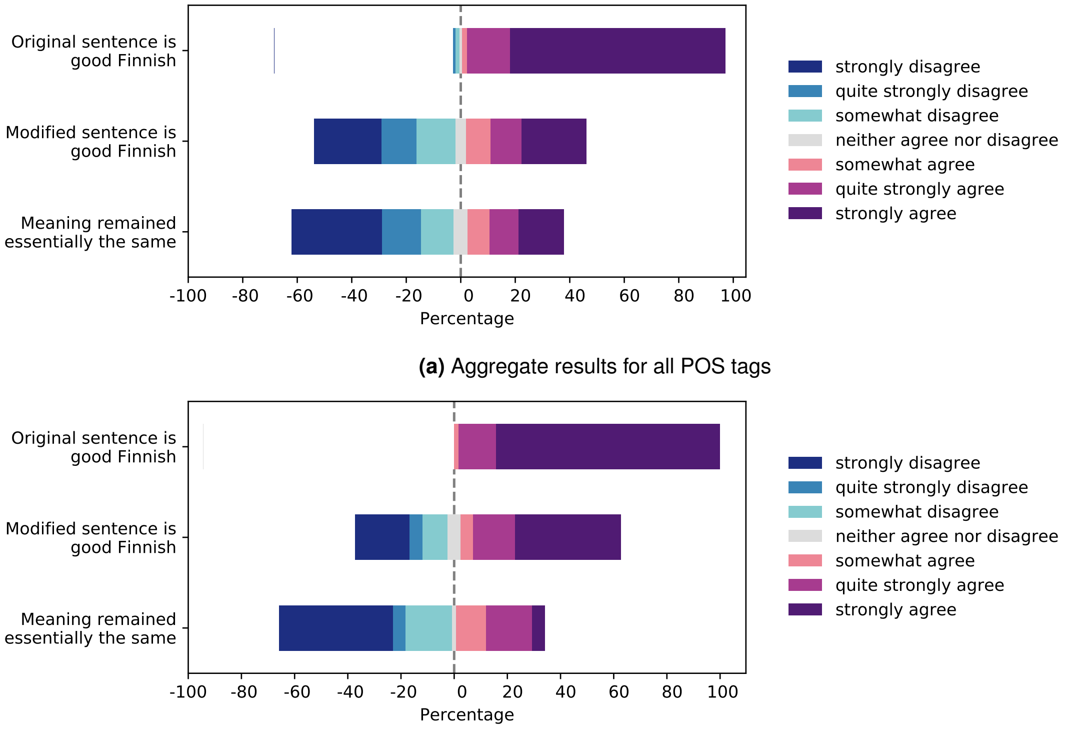


Figure 8: Results for questions 4 and 5 from applying the high-resource relexicalization approach to English.

The high-resource variant of the relexicalization approach (Figures 7 and 8), shows results that indicate the quality of the language stays the same with perhaps a slight decrease in the very low scores. This might be a consequence of the fact that in relexicalization, the number of tokens in the sentences stays constant, whereas in lexicalization new words are introduced. In any case, even a constant quality must be considered a success as it indicates the method having successfully introduced variance into the language without compromising the quality. We also see that the sentences seem to not have been changed in meaning, which is similarly positive. Like above, we also note that improving the filtering further might significantly improve the results, as the results for Q4 and Q5 (Figure 8) indicate both many unsuitable ‘approved’ words and suitable ‘disapproved’ words.

The low-resource relexicalization variant, the results shown in Figures 9 and 10, show that, similar to the low-resource lexicalization results above, the modifications did compromise the quality of the resulting sentences. Like above, change in sentence meaning was also observed. The results here were obtained using the lemmatization approach described in Section 5.2 and post-hoc analysis indicates that errors in the morphological analysis results in many errors, with a technically correct but contextually incorrect morphological analysis being identified as the most likely root cause. This, together with the results for Q4 and Q5 shown in Figure 10, indicates that significantly stronger results could be obtained by further refining the morphological analysis process and/or the word filtering. Like in the low-resource lexicalization case, the results for the best POS tag (in this case, nouns), were better than the general results. At the same time, even in this best case a significant drop in quality and change in meaning are observable. We remind the reader that additional noise is added to the method as a consequence of the round trip made to a high-resource language using the cross-lingual word embeddings, which likely explains at least some of the performance difference.

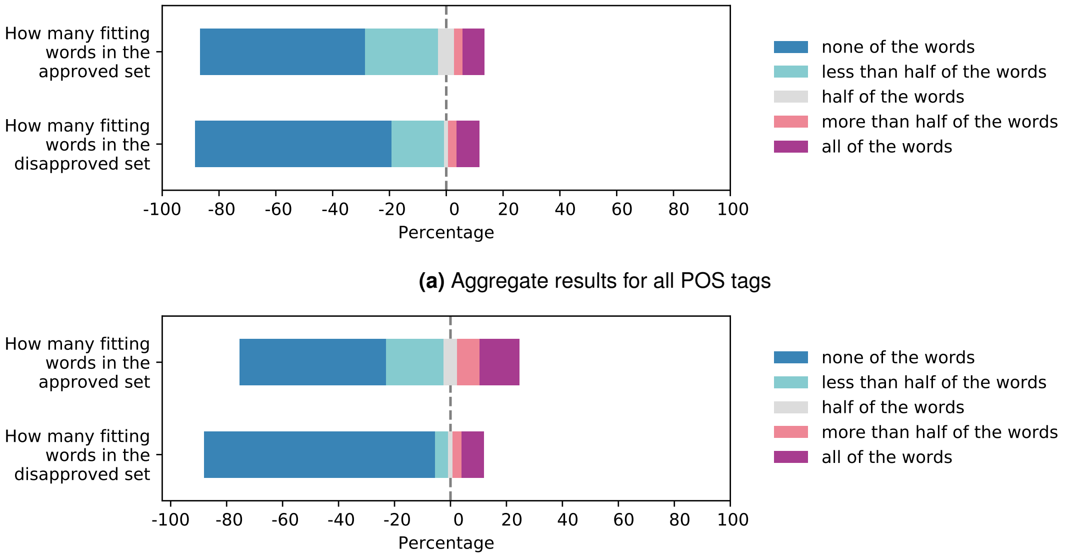
Examples of the relexicalization methods output are show in Figure 2. Here, the low-scoring sentence (c) can not be made grammatical and meaningful quite as easily. The error is probably related to the polysemy of the word ‘*vanhempien*’, which translates as both ‘*older*’ and ‘*parents*’. The generated replacement token seems to, in a way, the semantic-grammatical average of these two.



(a) Aggregate results for all POS tags

(b) Results for best POS tag: nouns

Figure 9: Results for questions 1–3 from applying the low-resource relexicalization approach to Finnish, with English as the high-resource language. Figure (a) presents the aggregate results over the various parts of speech which were trialed, where Figure (b) presents the results for what was identified as the best POS in post hoc analysis.



(a) Aggregate results for all POS tags

(b) Results for best POS tag: nouns

Figure 10: Results for questions 4 and 5 from applying the low-resource relexicalization approach to Finnish, with English as the high-resource language. Figure (a) presents the aggregate results over the various parts of speech which were trialed, where Figure (b) presents the results for what was identified as the best POS in post hoc analysis.

a) In 2017, 65 year old or older females did ~~earn~~ bring in mean net income of 27871 €.

b) Kyproksella vuonna 2017 ~~kotitaloudet~~ perheet maksoivat terveydenhuollon menoistaan itse 44.64 %.

c) Ranskassa vuonna 2017 75-vuotiaiden ja ~~vanhempien~~ *vanhuksien naisten tulojen keskiarvo oli 25770 €.

Figure 11: Example sentences produced using the relexicalization method in English (a) and Finnish (b, c). The section modified by relexicalization is denoted by underlining, with the original seed word struck over and the replacement word shown without strikethrough. Sentence (b) translates as '*In Cyprus in 2017, ~~households~~ families paid 44.64 % of their health care expenses themselves.*', and was scored well by the judges. Sentence (c) was scores low and contains an ungrammatical token denoted by *. It translates roughly as '*In France in 2017 the mean income of females aged 75 or ~~older~~ *elderly was 25770 €.*'

Overall, the results indicate that the suggested approaches are likely immediately useful for high-resource languages. We also interpret the results as indicating that modifying the filtering methods from pure POS-filters could also potentially improve the results noticeably. For the low-resource scenarios, the results indicate that the method is not immediately applicable at least in all cases. At the same time, the results indicate that the approach could be very useful with further fine-tuning. In interpreting these results, it is important to note that Finnish might present unique difficulties for this approach given its significant morphological complexity. In addition, as noted above, the results reported here were obtained without any fine-tuning of the word embeddings.

6 Potential of text readability measures in language generation

In the work by JSI and UL, partly related to EMBEDDIA, we have focused on automated readability assessment methods. We tested and adapted standard readability scores for Slovene text (Škvorc, Krek, Pollak, Špela Arhar-Holdt, & Robnik-Šikonja, 2019) and presented a set of novel approaches for determining readability of documents using deep neural networks (Martinc, Pollak, & Robnik-Šikonja, 2019). The main contributions of the papers are as follows:

- **Assessment of readability measures on Slovene texts**: (paper (Škvorc et al., 2019) in Appendix C). In this work, we adapt and test the readability measures, designed originally for English texts, to Slovene. We test ten well-known readability formulas and eight additional readability criteria on five types of texts: children's magazines, general magazines, daily newspapers, technical magazines, and transcriptions of national assembly sessions. As these groups of texts target different audiences, we assume that the differences in writing styles should be reflected in their readability scores.
- **A novel approach to readability measurement using RSRS score based on deep neural network based language models**: this approach is unsupervised and requires no labeled training set but only a collection of texts from the given domain. We demonstrate that the proposed approach is capable of contextualizing the readability because of the trainable nature of neural networks, and that it is transferable across different languages. We propose a new measure of readability, RSRS (ranked sentence readability score), with good correlation with true readability scores. For more details see the paper Martinc et al. (2019), in Appendix B.
- **A novel neural readability classification method**: we experiment how different neural architectures

with automatized feature generation can be used for readability classification and compare their performance to standard classification approaches, which rely on hand crafted features. Three distinct branches of neural architectures – recurrent neural networks (RNN), hierarchical attention networks (HAN), and transfer learning techniques – are tested on four gold standard readability corpora with excellent results.

In future, we will investigate how to incorporate this research into the news generation workflow. For example, assigning different levels of fluency/readability to the generated news text might be beneficial for domain-specific and more personalized news generation systems. First, readability scores might be used for adapting the text to different news genre (for example tabloid vs. political news), or personalising the production in terms of adaptation to different public (such as to non-native readers with different levels of language proficiency). Next, highlighting passages with very low readability might benefit the revision of automatically or manually generated text by editors or journalists. Finally, we want to integrate the readability measures directly into the text generation production. This can be seen as integration into the relexicalization phase described in Section 5.2. For example, when selecting different replacement words in template generation or template filling, we can compute the text readability, either by using standard readability measures, which can be adapted to all EMBEDDIA languages, as we did it for Slovene; by using our proposed RSDS score; or by adapting the readability classification task to various news production settings. In addition, the perplexity of a language models, which in our paper (Martinc et al., 2019) did not perform well for readability assessment, proved to work for fluency assessment (see (Liu Jr et al., 2020)). Perplexity score could also be used in headline generation approaches, with an approach similar to our work in slogan generation (Repar, Martinc, Znidarsic, & Pollak, 2018).

7 Associated outputs

The work described in this deliverable has resulted in the following resources:

Description	URL	Availability
Source code used in (re)lexicalization trials	`https://github.com/EMBEDDIA/nlg-manipulator`	To become public*

In addition, the work described in this deliverable is associated with the news generation case studies conducted in work package WP5 (see D5.2):

Description	URL	Availability
EuroStat news generation system (source code)	`https://github.com/EMBEDDIA/eurostat-nlg`	To become public*
COVID-19 news generation system (source code)	`https://github.com/EMBEDDIA/covid-nlg`	To become public*

* Resources marked here as “To become public” are available only within the consortium while under development and/or associated with work yet to be published. They will be released publicly and with a suitable open source license when the associated work is completed and published.

Parts of this work are also described in detail in the following publications (only partly related to EMBEDDIA), which are attached to this deliverable as appendices:

Citation	Status	Appendix
Vintar, Š., Simeunovic, L. G., Martinc, M., Pollak, S., & Stepisnik, U. (2020). Mining semantic relations from comparable corpora through intersections of word embeddings. In Proceedings of the 13th workshop on building and using comparable corpora (pp. 29–34).	Published	Appendix A
Martinc, M., Pollak, S., Robnik-Šikonja, M. Supervised and unsupervised neural approaches to text readability. `https://arxiv.org/pdf/1907.11779.pdf`	Submitted	Appendix B
Škvorc, T., Krek, S., Pollak, S., Arhar-Holdt, Š., Robnik-Šikonja, M.: Predicting Slovene Text Complexity Using Readability Measures. Contributions to Contemporary History (Digital Humanities and Language Technologies) 59 (1). `https://ojs.inz.si/pnz/article/view/323`	Published	Appendix C

8 Conclusions and further work

We have presented a modular pipeline architecture for language generation that enables multilingual natural language generation in either a fully rule-based manner, or with the incorporation of hybrid techniques, depending on the specific requirements imposed by the generation setting. While evaluating the ‘success’ of a general approach is non-trivial, promising results have already been achieved by using the approach in the initial versions of two distinct news generation systems described in Deliverable D5.2, which successfully applied the approach described herein to two distinct types of news generation tasks. Our analysis of even the initial implementations described in D5.2 indicated that the architecture succeeded in meeting the requirements identified in requirement analysis.

We have also demonstrated two methods for lexicalizing and relexicalizing generated sentences for added fluency and variability using contextual word embeddings. These methods showed promising initial results in human evaluations by producing more varied language without sacrificing the correctness of the output. We also described two variants of the algorithms that take advantage of cross-lingual word embeddings to allow the use of linguistic resources from high-resource languages to be used with low-resource languages. While the human evaluations here were not as good as those observed in the case of high-resource languages, they still seem promising enough to warrant future investigation both within and outside the EMBEDDIA research project.

In the future, we intend to investigate how to improve the (re)lexicalization methods and integrate them into the news generation systems described in Deliverable D5.2. Similarly, we will continue refining the language generation approach described in Section 4 within Work Packages WP5 and WP3.

References

Belz, A., & Kow, E. (2010). Extracting parallel fragments from comparable corpora for data-to-text generation. In *Proceedings of the 6th international natural language generation conference* (pp. 167–171).

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1724–1734).

Coch, J. (1998). Multimeteo: multilingual production of weather forecasts. *ELRA Newsletter*, *3*(2).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American* *chapter of the Association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186).

Dušek, O., Howcroft, D. M., & Rieser, V. (2019). Semantic noise matters for neural natural language generation. In *Proceedings of the 12th international conference on natural language generation* (pp. 421–426).

Dušek, O., Novikova, J., & Rieser, V. (2018). Findings of the E2E NLG challenge. *arXiv preprint arXiv:1810.01170*.

Ferreira, T. C., van der Lee, C., van Miltenburg, E., & Krahmer, E. (2019). Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. *arXiv preprint arXiv:1908.09022*.

Gatt, A., & Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, *61*, 65–170.

Gkatzia, D. (2016). Content selection in data-to-text systems: A survey. *arXiv preprint*. (Available at `https://arxiv.org/abs/1610.08375`)

Goldberg, E., Driedger, N., & Kittredge, R. I. (1994). Using natural-language processing to produce weather forecasts. *IEEE Expert*, *9*(2), 45–53.

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41–58). Brill.

Karlsson, F. (1996). *The word-forms of the finnish noun kauppa 'shop'.* (Available online: `http://www.ling.helsinki.fi/ fkarlsso/genkau2.html`)

Leppänen, L., Munezero, M., Granroth-Wilding, M., & Toivonen, H. (2017). Data-driven news generation for automated journalism. In *Proceedings of the 10th international conference on natural language generation* (pp. 188–197).

Linden, C.-G. (2017). Decades of automation in the newsroom: Why are there still so many jobs in journalism? *Digital Journalism*, *5*(2), 123–140.

Liu, C.-W., Lowe, R., Serban, I. V., Noseworthy, M., Charlin, L., & Pineau, J. (2016). How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

Liu Jr, Y., et al. (2020). Assessing text readability and quality with language models.

Martinc, M., Pollak, S., & Robnik-Šikonja, M. (2019). Supervised and unsupervised neural approaches to text readability. *arXiv preprint arXiv:1907.11779*.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, *38*(11), 39–41.

Moryossef, A., Goldberg, Y., & Dagan, I. (2019). Step-by-step: Separating planning from realization in neural data-to-text generation. *arXiv preprint arXiv:1904.03396*.

Nie, F., Yao, J.-G., Wang, J., Pan, R., & Lin, C.-Y. (2019). A simple recipe towards reducing hallucination in neural surface realisation. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 2673–2679).

Oh, A., & Rudnicky, A. (2000). Stochastic language generation for spoken dialogue systems. In *Anlp-naacl 2000 workshop: Conversational systems.*

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 conference of the North American chapter of the Association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 2227–2237).

Portet, F., Reiter, E., Gatt, A., Hunter, J., Sripada, S., Freer, Y., & Sykes, C. (2009). Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, *173*(7-8), 789–816.

Puduppully, R., Dong, L., & Lapata, M. (2019). Data-to-text generation with content selection and planning. In *Proc. 33rd aaai conference on artificial intelligence.*

Reiter, E. (2007). An architecture for data-to-text systems. In *Proceedings of the eleventh european workshop on natural language generation* (pp. 97–104).

Reiter, E. (2018). *Hallucination in neural nlg.* `https://ehudreiter.com/2018/11/12/hallucination-in-neural-nlg/`. (Accessed: 2020-03-02)

Reiter, E., & Belz, A. (2009). An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, *35*(4), 529–558.

Reiter, E., & Dale, R. (2000). *Building natural language generation systems*.

Reiter, E., Sripada, S., Hunter, J., Yu, J., & Davy, I. (2005). Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, *167*(1-2), 137–169.

Repar, A., Martinc, M., Znidarsic, M., & Pollak, S. (2018). Bislon: Bisociative slogan generation based on stylistic literary devices. In *Iccc.*

Sirén-Heikel, S., Leppänen, L., Lindén, C.-G., & Bäck, A. (2019). Unboxing news automation: Exploring imagined affordances of automation in news journalism. *Nordic Journal of Media Studies*, *1*(1), 47–66.

Vintar, Š., Simeunović, L. G., Martinc, M., Pollak, S., & Stepišnik, U. (2020). Mining semantic relations from comparable corpora through intersections of word embeddings. In *Proceedings of the 13th workshop on building and using comparable corpora* (pp. 29–34).

Škvorc, T., Krek, S., Pollak, S., Špela Arhar-Holdt, & Robnik-Šikonja, M. (2019). Predicting slovene text complexity using readability measures. *Contributions to Contemporary History: Digital Humanities and Language Technologies*, *59/1*.

Wen, T.-H., Gasic, M., Mrkšić, N., Su, P.-H., Vandyke, D., & Young, S. (2015). Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1711–1721).

Yleisradio. (2018). *Avoin voitto.* `https://github.com/Yleisradio/avoin-voitto`. GitHub.

You, Q., Jin, H., Wang, Z., Fang, C., & Luo, J. (2016). Image captioning with semantic attention. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4651–4659).

Appendix A: Mining Semantic Relations from Comparable Corpora through Intersections of Word Embeddings

Proceedings of the 13th Workshop on Building and Using Comparable Corpora, pages 29–34
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Mining Semantic Relations from Comparable Corpora through Intersections of Word Embeddings

Špela Vintar[1], Larisa Grčić Simeunović[2], Matej Martinc[3], Senja Pollak[4], Uroš Stepišnik[5]
University of Ljubljana[1,5], University of Zadar[2], Jožef Stefan Institute[3,4]
Aškerčeva 2, SI-1000 Ljubljana, M. Pavlinovića 1, HR-23000 Zadar, Jamova cesta 39, SI-1000 Ljubljana
spela.vintar@ff.uni-lj.si, lgrcic@unizd.hr, matej.martinc@ijs.si, senja.pollak@ijs.si, uros.stepisnik@ff.uni-lj.si

Abstract

We report an experiment aimed at extracting words expressing a specific semantic relation using intersections of word embeddings. In a multilingual frame-based domain model, specific features of a concept are typically described through a set of non-arbitrary semantic relations. In karstology, our domain of choice which we are exploring though a comparable corpus in English and Croatian, karst phenomena such as landforms are usually described through their FORM, LOCATION, CAUSE, FUNCTION and COMPOSITION. We propose an approach to mine words pertaining to each of these relations by using a small number of seed adjectives, for which we retrieve closest words using word embeddings and then use intersections of these neighbourhoods to refine our search. Such cross-language expansion of semantically-rich vocabulary is a valuable aid in improving the coverage of a multilingual knowledge base, but also in exploring differences between languages in their respective conceptualisations of the domain.

Keywords : semantic relations, word embeddings, comparable corpus, karstology, frame-based terminology

1. Introduction

The frame-based approach in terminology (FBT; Faber, 2012; Faber, 2015; Faber & Cabezas-García, 2019) has brought the notion that specialised knowledge can be modelled through conceptual frames which simulate the cognitive patterns in our minds. According to Faber (2012), "[a] frame is thus as an organized package of knowledge that humans retrieve from long-term memory to make sense of the world." Two of the most significant practical contributions of FBT are on the one hand the consolidation between the conceptual and the textual level of domain representation by using specialised corpora for the induction of frames or event templates, and on the other hand the realisation that such frames and templates are not universal but contextually, culturally and linguistically bound.

On a more practical level, the frame-based approach to domain modelling fosters a dynamic and process-oriented view of the concepts, actions, properties and events leading to a deeper understanding of the domain. This is particularly relevant for a domain such as karstology where karst landscapes and landforms are the result of complex and prolonged natural processes occurring in specific environments and under specific sets of conditions.

The broader context for this research is the TermFrame project which employs and extends the frame-based approach to build a visual knowledge base for the domain of karstology in three languages, English, Slovene and Croatian; as well as explores new methods of knowledge extraction from specialized texts (Vintar et al., 2019, Miljkovic et al., 2019, Pollak et al. 2019).

The domain of karstology is conceptualized in terms of events where natural or human agents initiate actions or processes which affect patients in specific ways and thus result in various karst features. In order to explore typical conceptual frames in karstology we devised a domain-specific concept hierarchy of semantic categories, and each category can be described by a set of relations which reveal its typical features. For example, the category of *surface landforms* is typically described by relations that express form, size, location and cause while concepts from the category of *hydrological landforms* are usually defined by the relations cause, location and function.

When building a multilingual knowledge base, identifying such relations is important from the perspective of organising knowledge and ensuring maximum coverage of the domain. For example, COMPOSITION in terms of geological structure plays a crucial role in karstology because karst phenomena can only develop on soluble rocks. It is therefore extremely useful if we can access the entire inventory of expressions denoting COMPOSITION in our corpus, and also compare them between languages as this gives important clues about the domain itself, e.g. the prominence of certain minerals in different geographical regions.

In this research we propose a method to extract expressions pertaining to a specific semantic relation from a comparable English and Croatian corpus by providing a limited number of seed words for each language and relation, then using word embeddings to identify words belonging to same relation class. The seed words in our study are limited to adjectives because of their combinatorial potential within multi-word terms and the observation that semantic relations are frequently expressed through adjectives.

2. Related work

One of the aims of this study is to leverage word embeddings and a set of seed adjectives expressing semantic relations in order to extract additional adjectives that express the same semantic relation/attribute. This is in essence a set expansion task and previous research on a related subject was conducted by Diaz et al. (2016), who showed that embeddings can be employed for query expansion on domain specific texts. The research concludes that due to strong language use variation in specialized corpora, domain specific embeddings (trained locally on a small specialized corpora) outperform non-topic specific general embeddings trained on a much larger general corpus. A very similar approach for set expansion in the domain of karstology was employed by Pollak et al. (2019) for the purposes of extending terminology.

Previous authors (Duran Muñoz, 2019, Bhat, 1994, Wierzbicka, 1986, Fellbaum et al., 1993, L'Homme, 2002) have already examined the role of adjectives in specialised languages and confirmed their importance in expressing key properties of specialized concepts as well as appearing as parts of multi-word terms. A particularly relevant analysis of semantic relations in complex nominals was performed by Cabezas-García and León-Araúz (2018), who use knowledge patterns and verb paraphrases to construct a frame-based model of semantic categories and the semantic relations occurring between them. They show that a particular combinatorial pattern established for a set of nouns can be extrapolated to the entire semantic category and potentially used for relation induction.

We are also aware of several studies describing the semantic representation of adjectives in ontologies for other domains, e.g. legal (Bertoldi and Chisman, 2007), environment (Campos Alonso and Castells Torner, 2010), plant morphology (Pitkanen-Heikkila, 2015) and waste management (Altmanova et al., 2018).

3. Karstology and the TermFrame Corpus

Karstology is the study of karst, a type of landscape developing on soluble rocks such as limestone, marble or gypsum. Its most prominent features include caves, various types of relief depressions, conical hills, springs, ponors and similar. It is an interdisciplinary domain partly overlapping with surface and subsurface geomorphology, geology, hydrology and other fields.

For the purposes of our research, we used the English and Croatian parts of the TermFrame corpus, which otherwise also contains Slovene as the third language. The comparable corpus contains relevant contemporary works on karstology and is representative in terms of the domain and text types included. It comprises scientific papers, books, articles, doctoral and master's theses, glossaries and textbooks. Table 1 gives basic information about the corpus.

	English	Croatian
Tokens	2,721,042	1,229,368
Words	2,195,982	969,735
Sentences	97,187	53,017
Documents	57	43

Table 1: Corpus information

4. Methods

4.1 Framing karstology

The TermFrame project models the karstology domain using a hierarchy of semantic categories and a set of relations which allow us to describe and model karst events (Vintar et al., 2019). According to the geomorphologic analytical approach (Pavlopoulos et al., 2009), the relations describe different aspects of concepts, such as spatial distribution (HAS_LOCATION; HAS_POSITION), morphography (HAS_FORM; CONTAINS), morphometry (HAS_SIZE), morphostructure (COMPOSED_OF), morphogenesis (HAS_CAUSE), morphodynamics (AFFECTS; HAS_RESULT; HAS_FUNCTION), and morphochronology (OCCURS_IN_TIME). Additional relations were applied for general properties (HAS_ATTRIBUTE; DEFINED_AS), and for research methods (STUDIES; MEASURES).

The research described here focuses on the 5 relations which occur most frequently in the definitions of karst landforms and processes, and they also govern the formation of multi-word terms as illustrated by examples below.

underground cave ⇒ LOCATION (cave) = underground

fluvial sediment ⇒ CAUSE (sediment)=fluvial

enclosed depression ⇒ FORM (depression)= enclosed

gypsum karst ⇒ COMPOSITION (karst)=gypsum

soluble rock ⇒ FUNCTION (rock)=soluble

We thus examined the contexts expressing the selected relations in the TermFrame corpus of annotated definitions (Vintar et al., 2019). From these contexts we obtained lists of seed adjectives for each relation and both languages, which were validated by a domain expert:

LOCATION
English*: coastal, littoral, sublittoral, submarine, oceanic, subsurface, subterranean, subterraneous, subaerial, underground, aquatic, subaqueous, internal, subglacial, epigenic, phreatic, vadose, epiphreatic*

Croatian*: obalni, litoralan, priobalni, podmorski, oceanski, podzeman, freatski, vadozan, podvodan, dolinski, špiljski, epifreatski*

CAUSE
English: fluvial, allogenic, tectonic, erosional, alluvial, volcanic, lacustrine, solutional, aeolian, periglacial, anthropogenic

Croatian: fluvijalni, alogeni, tektonski, erozijski, aluvijalan, vulkanski, lakustrijski, eolski, periglacijalni, antropogeni

FORM
English: *polygonal, vertical, dendritic, shallow, enclosed, elongated, flat, steep, cavernicolous, detrital*

Croatian: vertikalan, ravnocrtan, strm, kavernozan, horizontalan, mrežast, longitudinalan, kružan, razgranat, ulegnut, uravnjen

COMPOSITION
English: *carbonate, limestone, dolomitic, sedimentary, sulfate, calcareous, carboniferous, silicate, sulfuric, diagenetic, siliceous, clay, volcanoclastic*

Croatian: karbonatni, vapnenački, dolomitski, sedimentan, sulfatni, kalcitan, karbonski, sulfatni, glinovit, sedreni, stijenski,klastičan,sedreni

FUNCTION
English: impermeable, permeable, solutional, hydrothermal, speleological, geological, soluble, porous, depositional, regressive, undersaturated

Croatian: nepropustan, propustan, speleološki, geološki, topiv, porozan, taložan, urušan

4.2 Word embeddings

Our initial assumption was that the word embeddings of a set of adjectives expressing a specific semantic relation, such as CAUSE, FORM or COMPOSITION, share a certain semantic component which can be used to extract other adjectives expressing the same relation.

To test this assumption, we first train FastText embeddings (Bojanowski et al., 2017) on the English and the Croatian part of the TermFrame corpus respectively (see Section 3). Embeddings were calculated for all the words that appear in the corpus at least three times and we use a skip-gram model with an embedding dimension of 100. For each seed adjective expressing a specific semantic relation, we use embeddings to find a set of 100 closest words according to the cosine distance. In order to find words of similar semantic provenance that express a specific semantic relation, in the next step we calculate all non-empty intersections between these sets of 100 closest words for all possible subsets of a set of adjectives for each relation. These subsets range in size from 10 to 2, since 10 is the largest subset of seed adjectives for a relation, for which a non-empty intersection was returned. All words found in these intersections are retained as candidate words that express a specific relation and are used in manual evaluation (see Section 5). For example, (see examples (1) and (2) below), the intersection of the closest embeddings for a subset of 5 English input words for LOCATION (*coastal, littoral, oceanic, submarine, subterranean*) yields the single word *nonmarine* as intersection, while the intersection for the subset of 3 Croatian input words for FORM (*horizontalan, kružan, vertikalan*) yields 8 words in the intersection:

(1) SIZE: 5
SUBSET: coastal, littoral, oceanic, submarine, subterranean INTERSECTION: nonmarine

(2) SIZE: 3 SUBSET: horizontalan, kružan, vertikalan INTERSECTION: okomito, sjecište, vodoravan, inverzan, okomit, nepravilan, presjecište, konveksan

5. Results and Discussion

Intersections were computed for subsets of input words ranging from maximum 10 to 2 words, whereby most intersections were empty for larger subsets and only started yielding results from size 7 downwards (see Table 2).

Our first observation is that both in English and Croatian a large majority of extracted words are adjectives and other words functioning as premodifiers in multi-word terms, thus illustrating that the embeddings capture also syntactic properties.

Since the overall goal of the experiment is to extract words pertaining to the same semantic relation, we first report the total number of extracted words and the number of correctly predicted ones, i.e. belonging to the same semantic class as the input words (Table 2).

	location		function		form		composition		cause	
	en	cr	en	cr	en	cr	en	cr	en	cr
N	357	228	147	152	164	152	293	244	183	181
C	118	88	68	43	108	97	184	197	88	132
P	0.33	0.39	0.46	0.28	0.66	0.64	0.63	0.80	0.48	0.73

Table 2: Precision per semantic relation and language (N = number of extracted words, C = correct, P = precision (C/N))

A quick glance at Table 2 shows that the numbers of extracted words are slightly lower for Croatian, which is possibly due to the difference in the size of corpora, but the overall lowest and highest precisions are also found for Croatian candidates. Next we observe large differences between individual semantic relations, both in terms of precision of prediction and the yield, but relatively similar performance across both languages. The largest number of correctly extracted candidates is achieved for COMPOSITION, where an input of only 13 words allows us to extract 184 English and 197 Croatian expressions for geological or chemical composition, e.g. *lithoclast, calcitic, azurite, loessic, gneiss, chalky, magmatic, pyrite, framestone, siliclastic* and *kalkarenit, laporovit, škriljac, glinenac, piroksenit, fliški* etc. Many of the extracted expressions are highly specialised and occur in the corpus with a very low frequency, yet their membership in the semantic class could still be correctly predicted.

On the other hand, the LOCATION relation is more difficult to capture because it may refer to the position of an entity within the karst system, its position relative to some other entity or its position relative to the land or sea. The retrieved words include many geographical names, e.g. *Baltic*, *Bahamian;kvarnerski, mosorski*, which we do not count as positives for the simple reason that our annotation scheme uses a different semantic relation (HAS_POSITION) for toponyms.

Next, we measure the precision of the predicted relation for each intersection, and we report average precision for each subset size and each language (see Table 3 and Table 4). We use precision@M denoting the number of true predictions divided by the number of all words in the intersection, and precision@5 where the size of the intersection is fixed to 5 words. In this case, a perfect precision is not possible for intersections containing less than 5 words and intersections containing more than 5 words are truncated. For the example (1) above, precision@M = 1 and precision@5 = 0.2.

As mentioned before, most intersections for larger subsets (English 8-10 input words, Croatian 7-10 input words) were empty, except for COMPOSITION in English. This would indicate that the most suitable subset size ranges from 2 to 6 input words. In English, poorest results were obtained for FUNCTION, where the intersections of subsets 4-6 contained only a single word (*sluggish*), which expresses manner of (water) movement but not function. Results for FORM, COMPOSITION and CAUSE were however promising in that they yielded highly accurate predictions, e.g. *zigzag, honeycomb, steep, curvilinear, elliptical, coalescent, sharp, semicircular, asymmetric, sinusoidal, pinnacled, undulating* for FORM and *compressional, geogenic, preglacial, bioclastic, erosional, disolutional, orogenic, tensional* etc. for CAUSE.

	location		function		form		composition		cause	
subset size	p@M	p@5	p@M	p@5	p@M	p@5	p@M	p@5	p@M	p@5
10							1	0.20		
9							1	0.20		
8							1	0.21		
7	0	0					0.99	0.24	1	0.20
6	0.36	0.07	0	0	1	0.2	0.98	0.28	0.78	0.16
5	0.45	0.13	0	0	1	0.22	0.95	0.35	0.65	0.16
4	0.45	0.17	0.01	0	1	0.31	0.91	0.44	0.60	0.20
3	0.42	0.22	0.10	0.03	0.94	0.47	0.85	0.53	0.60	0.30
2	0.37	0.29	0.26	0.13	0.70	0.55	0.75	0.59	0.56	0.39

Table 3: Precision of English predicted words per subset size

	location		function		form		composition		cause	
subset size	p@M	p@5	p@M	p@5	p@M	p@5	p@M	p@5	p@M	p@5
6	0	0								
5	0	0	0.33	0.20	1	0.20	0	0	0.50	0.10
4	0.10	0.05	0.33	0.28	0.92	0.20	0.69	0.20	0.53	0.15
3	0.28	0.16	0.32	0.30	0.78	0.28	0.79	0.35	0.65	0.27
2	0.33	0.30	0.32	0.20	0.72	0.49	0.79	0.62	0.72	0.55

Table 4: Precision of Croatian predicted words per subset size

FUNCTION also had the lowest yield of meaningful expressions in Croatian, with only one non-empty intersection for subset 5, but on the other hand the entire range of karst-related studies was retrieved by intersecting *geološki* and *speleološki (3)*:

(3) SIZE: 2
SUBSET: geološki, speleološki
INTERSECTION: arheološki, biospeleološki, geomofološki, tipološki, geoekološki, biološki, mitološki, kršološki, ontološki, geoekološka, aerološki, fiziološki, paleokrški, speleomorfološki, drološki, geokronološki, etnološki, paleontološki, filološki

Results for English also show a positive linear correlation between the subset size and precision@M (especially for the relations FORM, COMPOSITION AND CAUSE), and a negative linear correlation between the subset size and precision@5. This phenomenon can be explained with the fact that at large subset sizes there are less than five words in the intersection which has a negative impact on precision@5, but as the few extracted examples are likely to be correct, it has a positive impact on precision@M. On the other hand, at small subset sizes the number of words in the intersection will increase, which has a positive effect on precision@5 but also negatively affects precision@M, since the percentage of correctly retrieved words in the intersection decreases. The results for Croatian also show a strong negative linear correlation between the subset size and precision@5, while for precison@M the correlation somewhat varies between relations, ranging from being negative for LOCATION, CAUSE and COMPOSITION, to no correlation for FUNCTION, and to a positive correlation for the FORM relation. This means that for Croatian a larger subset size does not necessarily guarantee that a larger percentage of extracted examples will be correct.

To understand why relations perform differently in such an experimental setting we must consider their conceptual role within the frame-based domain model. It is clear that there can be an almost indefinite number of words used to describe the form of an entity in the karst landscape - think just of the multitude of underground forms found in caves. The embeddings thus successfully capture about one hundred expressions for FORM in each language, yet miss words like *ravničast, ponikvast, kavernozan, terasast, klifast, zaravnjen* etc. On the other hand, not all karst landforms have functions in the karstologic event, and the number of possible causes is also limited. For CAUSE, certain suffixes seem especially productive and allow us to extract relevant expressions – often cognates – on this basis: -genic/-gen, -genijski, -genski (*epigenic, geogenic, cryogenic, orogenic, biogenic, pathogenic, hypogenic, glacigenic, rheogenic / epigenijski, orogenski, egzogen, kemogen, zoogen, biogen, kriogen*); -glacial/-glacijalan (*preglacial, subglacial, fluvioglacial, englacial, proglacial, supraglacial / glacijalan, proglacijalan, interglacijalan, postglacijalan, fluvioglacijalan, periglacijalan*), -luvial/-luvijalan (*alluvial, eluvial, colluvial, pluvial, deluvial / iluvijalan, proluvijalan, delovijalan, diluvijalan, koluvijalan*).

In all experiments reported above we measure precision but not recall. To measure recall we would need to have a list of true positives for each relation, which could only be created manually by inspecting, for instance, all adjectives in the corpus and labelling them with relations, which has not been done as yet.

Finally, during evaluation we noted several ambiguous examples which in some contexts could refer to causes, while in others they denote composition, function or form. For Croatian, some overlap was found between the lists of expressions denoting COMPOSITION and FUNCTION (e.g. *vodopropusan* [permeable]), and for English between COMPOSITION and CAUSE (e.g. *magmatic, sediment, igneous*). Indeed such cases show that some relations are closer than others, and that specialised vocabulary is inherently multidimensional and context-dependent.

6. Conclusions

We explore semantic relations in a comparable English and Croatian corpus of karstology focusing on the adjectives and other premodifiers in multi-word terms. By assuming the frame-based domain model we identify groups of seed adjectives according to the semantic relation they express in the multi-word terms (e.g. FORM, LOCATION, FUNCTION), whereby the conceptual frame provides guidance as to which relations are expected for each concept category.

Against these background assumptions we attempt to extract attributes pertaining to the same relation using word embeddings computed on the two domain-specific corpora. We use subsets of seed adjectives as input and intersect their closest neighbours to extract candidate English and Croatian words.

Results are relatively similar across the two languages, but show high variability in precision between relations, with poor performance for the FUNCTION relation and slightly better for LOCATION. On the other hand, for the other three relations (COMPOSITION, FORM, CAUSE) results seem highly promising in that for both languages the intersections yield relevant candidates with high precision, despite the relatively small size of the domain-specific corpora. Our approach illustrates that word embeddings trained on small specialised corpora can be used to predict the semantic relations in a frame-based setting.

As future work we plan to explore the possibility of modelling karstological processes and events using analogies between semantically related pairs of concepts. It appears that the cognitive dimensions of frame-based knowledge modelling have interesting parallels within the spatial logic of word embeddings.

It is also possible to imagine a scenario where word embeddings and intersections of related words can be used to develop a frame-based model for a new domain, or more specifically to help discern the relations.

Another line of future work will consider cross-lingual query expansion, where we will try to extract adjectives expressing a specific relation in the target language by using only seed terms from the source language. In order to do this we would first need to align embeddings for both languages into a common vector space by using one of the existing methods, e.g., the one proposed in Conneau et. al (2017) that also employs FastText embeddings. Leveraging this procedure we would be able to expand the set of adjectives in a target language with terms that are not clearly associated with the target language seed terms but do however express the same relation.

7. Acknowledgements

This work is supported by the Slovenian Research Agency grant J6-9372 TermFrame: Terminology and knowledge frames across languages, 2018-2021 and by European Union's Horizon 2020 research and innovation programme under grant agreement No. 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media).

8. Bibliographical References

Altmanova, J., Grimaldi, C., Zollo, S. D. (2018). Le rôle des adjectifs dans la catégorisation des déchets. In F. Neveu, B. Harmegnies, L. Hriba et S. Prévost (Eds.), *SHS Web Conferences* 46, 6ème Congrès Mondial de Linguistique Française. Université de Mons, Belgique, pp. 1-15.

Bhat, D.N.S. (1994). The adjectival category: Criteria for differentiation and identification. Amsterdam: John Benjamins Publishing Co.

Bertoldi, A., Chishman, R.L. (2007). Improving Legal Ontologies through Semantic Representation of Adjectives. ICSC 2007, pp. 767-774.

Bojanowski, P. et al. (2017) Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5, pp: 135-146.

Cabezas-García, M., and León-Araúz, P. (2018). Towards the inference of semantic relations in complex nominals: A pilot study. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), pp. 2511-2518.

Campos, Alonso, A. and Castells, Torner S. (2010) Adjectives and collocations in specialized texts: lexicographical implications. In A. Dykstra, T. Schoonheim (Eds.), Proceedings of the XIV EURALEX International Congress. Leeuwarden/Ljouwert: Fryske Akademy - Afûk. pp. 872-881.

Conneau, A., Lample, G., Ranzato, M. A., Denoyer, L. and Jégou, H. (2017). Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Diaz, F., Bhaskar M., Craswell, N. (2016). Query expansion with locally-trained word embeddings. *arXiv preprint arXiv:1605.07891*

Durán-Muñoz, I. (2019/forthcoming). Adjectives and their Keyness. A Corpus-based Analysis in English Tourism. *Corpora*, 14 (3). Edinburgh University Press.

Faber, P. (Ed.), (2012). A Cognitive Linguistics View of Terminology and Specialized Language. Berlin and New York: Mouton De Gruyter.

Faber, P. (2015). Frames as a framework for terminology. In Kockaert, H.J., Steurs, F. (Eds) *Handbook of Terminology*. Vol 1. Amsterdam and Philadelphia: John Benjamins, pp. 14-33.

Faber, P., Cabezas-García, M. (2019) Specialized Knowledge Representation: From Terms to Frames. *Research in Language*, 17(2): 197-211.

Fellbaum, C., Gross, D., Miller, K. (1993). Adjectives in WordNet. In G. Miller et al. (Eds.) Five Papers on WordNet. Tehnical Report 43, Cognitive Science Laboratory, Princeton University, pp. 26-39.

L'Homme, M.C. (2002). What can Verbs and Adjectives Tell us about Terms? Paper presented at TKE 2002, Nancy, France.

Miljković, D., Kralj, J., Stepišnik, U., Pollak, S. (2019). Communities of related terms in a karst terminology co-occurrence network. *Proceedings of eLex19*, Sintra, Portugal. pp. 357-373.

Pavlopoulos, K., Evelpidou, N., Vassilopoulos, A. (2009). *Mapping Geomorphological Environments*. Springer, Berlin Heidelberg.

Pitkänen-Heikkilä, K. (2015) Adjectives as terms. *Terminology*, 21 (1):76-101

Pollak, S., Repar, A., Martinc, M., Podpečan, V. (2019). Karst exploration: extracting terms and definitions from karst domain corpus. *Proceedings of eLex19*, Sintra, Portugal. pp. 934-956.

Vintar, Š., Saksida, A., Vrtovec, K., Stepišnik, U. (2019). Modelling specialized knowledge with conceptual frames: the TermFrame approach to a structured visual domain representation. *Proceedings of eLex19*, Sintra, Portugal. pp. 305-318.

Wierzbicka A. (1986) What's in a noun? (Or: How do nouns differ in meaning from adjectives?) *Studies in Language* 10:353-389.

Appendix B: Supervised and unsupervised neural approaches to text readability

Supervised and unsupervised neural approaches to text readability

Matej Martinc.
E-mail:matej.martinc@ijs.si.
Jožef Stefan Institute, Ljubljana, Slovenia

Senja Pollak.
E-mail:senja.pollak@ijs.si.
Jožef Stefan Institute, Ljubljana, Slovenia
Usher Institute of Population Health Sciences, University of Edinburgh, UK

Marko Robnik-Šikonja.
E-mail:marko.robnik@fri.uni-lj.si
University of Ljubljana, Faculty of Computer and Information Science, Ljubljana, Slovenia

We present a set of novel neural supervised and unsupervised approaches for determining readability of documents. In the unsupervised setting, we leverage neural language models, while in the supervised setting three different neural architectures are tested in the classification setting. We show that the proposed neural unsupervised approach on average produces better results than traditional readability formulas and is transferable across languages. Employing neural classifiers, we outperform current state-of-the-art classification approaches to readability which rely on standard machine learning classifiers and extensive feature engineering. We tested several properties of the proposed approaches and showed their strengths and possibilities for improvements.

1. Introduction

Readability is concerned with the relation between a given text and the cognitive load of a reader to comprehend it. This complex relation is influenced by many factors, such as a degree of lexical and syntactic sophistication, discourse cohesion, and background knowledge (Crossley et al. 2017). In order to simplify the problem of measuring readability, traditional readability formulas focused only on the lexical and syntactic features by taking into an account various statistical factors, such as word length, sentence length, and word difficulty (Davison and Kantor 1982). These approaches have been criticized because of their reductionism and weak statistical bases (Crossley et al. 2017). Another problem is their objectivity and cultural transferability, since children from different environments master different concepts at different ages. For example, a word *television* is quite long and contains many syllables but is well-known to most young children who live in families with a television.

Newer approaches to measuring readability consider it as a classification task and build prediction models that predict human assigned readability scores based on a number of attributes (Schwarm and Ostendorf 2005; Vajjala and Meurers 2012; Petersen and Ostendorf 2009). These more sophisticated and adaptable approaches generally yield better results and are less exposed to critique but require additional external resources, such as labeled readability data sets, which are scarce. Another problem is transferability of these approaches between different corpora and languages, since little work has been done on multilingual, multi-genre, or even multi-corpora supervised approaches to readability prediction.

Recently, deep neural networks (Goodfellow, Bengio, and Courville 2016) have shown impressive performance on many language related tasks. In fact, they have achieved the state-of-the-art performance in all semantic tasks where sufficient amounts of data were available (Collobert et al. 2011; Zhang, Zhao, and LeCun 2015). Surprisingly, we are not aware of any work that would employ deep neural models for the task of determining readability. Even the most recent studies (Vajjala and Lucic 2018) still rely on hand-crafted features and standard classifiers, such as Support Vector Machines (SVM), when trying to determine text readability. Furthermore, language model features, which can be found in many of these classification approaches (Schwarm and Ostendorf 2005; Petersen and Ostendorf 2009; Vajjala and Meurers 2012; Xia, Kochmar, and Briscoe 2016), are generated with traditional n-gram language models, even though language modeling, which can be formally defined as predicting a probability distribution of words from the fixed size vocabulary V, for word $w_{\mathrm{t+1}}$, given the historical sequence $w_{\mathrm{1:t}} = [w_1, ..., w_t]$, has been drastically improved with the introduction of neural language models (Mikolov et al. 2011).

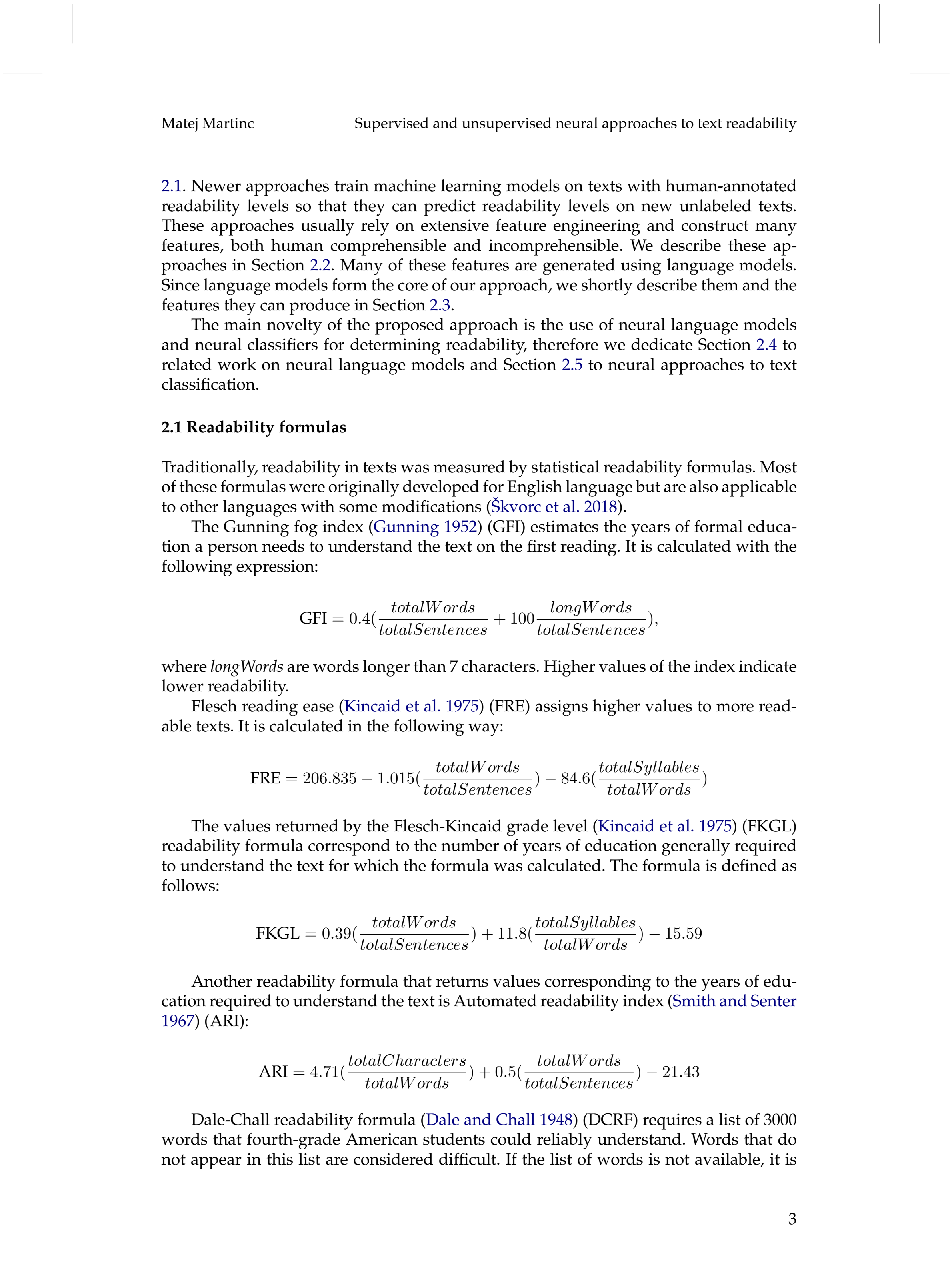
The aim of the present study is two-fold. First, we propose a novel approach to readability measurement based on deep neural network based language models that takes into account background knowledge and discourse cohesion, two readability indicators missing from the traditional readability formulas. This approach is unsupervised and requires no labeled training set but only a collection of texts from the given domain. We demonstrate that the proposed approach is capable of contextualizing the readability because of the trainable nature of neural networks, and that it is transferable across different languages. In this scope, we propose a new measure of readability, RSRS (ranked sentence readability score), with good correlation with true readability scores.

Second, we experiment how different neural architectures with automatized feature generation can be used for readability classification and compare their performance to standard classification approaches, which rely on hand crafted features. Three distinct branches of neural architectures – recurrent neural networks (RNN), hierarchical attention networks (HAN), and transfer learning techniques – are tested on four gold standard readability corpora with excellent results.

The paper is structured as follows. Section 2 addresses the related work on readability prediction and also covers more general topics related to our research, such as language modelling and neural text classification. Section 3 describes the datasets used in our experiments, while in Section 4 we present the methodology and results for the proposed unsupervised approach to readability prediction. The methodology and experimental results for the supervised approach are presented in Section 5. The conclusions and directions for further work are addressed in Section 6.

2. Background and related work

Approaches to automated measuring of readability try to find and assess factors that correlate well with human perception of readability. They can be divided into two groups. Traditional readability formulas try to construct a simple human comprehensible formula with a good correlation to what humans perceive as the degree of readability. They take into account various statistical factors, such as word length, sentence length, and word difficulty. We describe the most popular constructs in Section



2.1. Newer approaches train machine learning models on texts with human-annotated readability levels so that they can predict readability levels on new unlabeled texts. These approaches usually rely on extensive feature engineering and construct many features, both human comprehensible and incomprehensible. We describe these approaches in Section 2.2. Many of these features are generated using language models. Since language models form the core of our approach, we shortly describe them and the features they can produce in Section 2.3.

The main novelty of the proposed approach is the use of neural language models and neural classifiers for determining readability, therefore we dedicate Section 2.4 to related work on neural language models and Section 2.5 to neural approaches to text classification.

2.1 Readability formulas

Traditionally, readability in texts was measured by statistical readability formulas. Most of these formulas were originally developed for English language but are also applicable to other languages with some modifications (Škvorc et al. 2018).

The Gunning fog index (Gunning 1952) (GFI) estimates the years of formal education a person needs to understand the text on the first reading. It is calculated with the following expression:

$$\text{GFI} = 0.4(\frac{totalWords}{totalSentences} + 100\frac{longWords}{totalSentences}),$$

where *longWords* are words longer than 7 characters. Higher values of the index indicate lower readability.

Flesch reading ease (Kincaid et al. 1975) (FRE) assigns higher values to more readable texts. It is calculated in the following way:

$$\text{FRE} = 206.835 - 1.015(\frac{totalWords}{totalSentences}) - 84.6(\frac{totalSyllables}{totalWords})$$

The values returned by the Flesch-Kincaid grade level (Kincaid et al. 1975) (FKGL) readability formula correspond to the number of years of education generally required to understand the text for which the formula was calculated. The formula is defined as follows:

$$\text{FKGL} = 0.39(\frac{totalWords}{totalSentences}) + 11.8(\frac{totalSyllables}{totalWords}) - 15.59$$

Another readability formula that returns values corresponding to the years of education required to understand the text is Automated readability index (Smith and Senter 1967) (ARI):

$$\text{ARI} = 4.71(\frac{totalCharacters}{totalWords}) + 0.5(\frac{totalWords}{totalSentences}) - 21.43$$

Dale-Chall readability formula (Dale and Chall 1948) (DCRF) requires a list of 3000 words that fourth-grade American students could reliably understand. Words that do not appear in this list are considered difficult. If the list of words is not available, it is possible to use the GFI approach and consider all the words longer than 7 characters as difficult. The following expression is used in calculation:

$$\text{DCRF} = 0.1579(\frac{difficultWords}{totalWords} * 100) + 0.0496(\frac{totalWords}{totalSentences})$$

The SMOG grade (Simple Measure of Gobbledygook) (Mc Laughlin 1969) is a readability formula mostly used for checking health messages. Similar as FKGL and ARI, it roughly corresponds to the years of education needed to understand the text. It is calculated with the following expression:

$$\text{SMOG} = 1.0430\sqrt{numberOfPolysyllables\frac{30}{totalSentences}}3.1291,$$

where the *numberOfPolysyllables* is the number of words with three or more syllables.

All of the above mentioned readability measures were designed for the specific use on English texts. There are some rare attempts to adapt these formulas to other languages (Kandel and Moles 1958) or to create new formulas that could be used on languages other than English (Anderson 1981).

To show a cross-lingual potential of our approach, we address two languages in this study, English and Slovenian, a Slavic language with rich morphology and orders of magnitude less resources compared to English. For Slovenian, readability studies are scarce. Škvorc et al. (2018) researched how well the above readability formulas work on Slovenian text by trying to categorize text from three distinct sources: children's magazines, newspapers and magazines for adults, and transcriptions of sessions of the National Assembly of Slovenia. Results of this study indicate that formulas which consider the length of words and/or sentences work better than formulas which rely on word lists. They also noticed that simple indicators of readability, such as percentage of adjectives and average sentence length, also work quite well for Slovenian. To our knowledge, the only other study that employed readability formulas on Slovenian texts was done by Zwitter Vitez (2014). Here the readability formulas were used as features in the author recognition task.

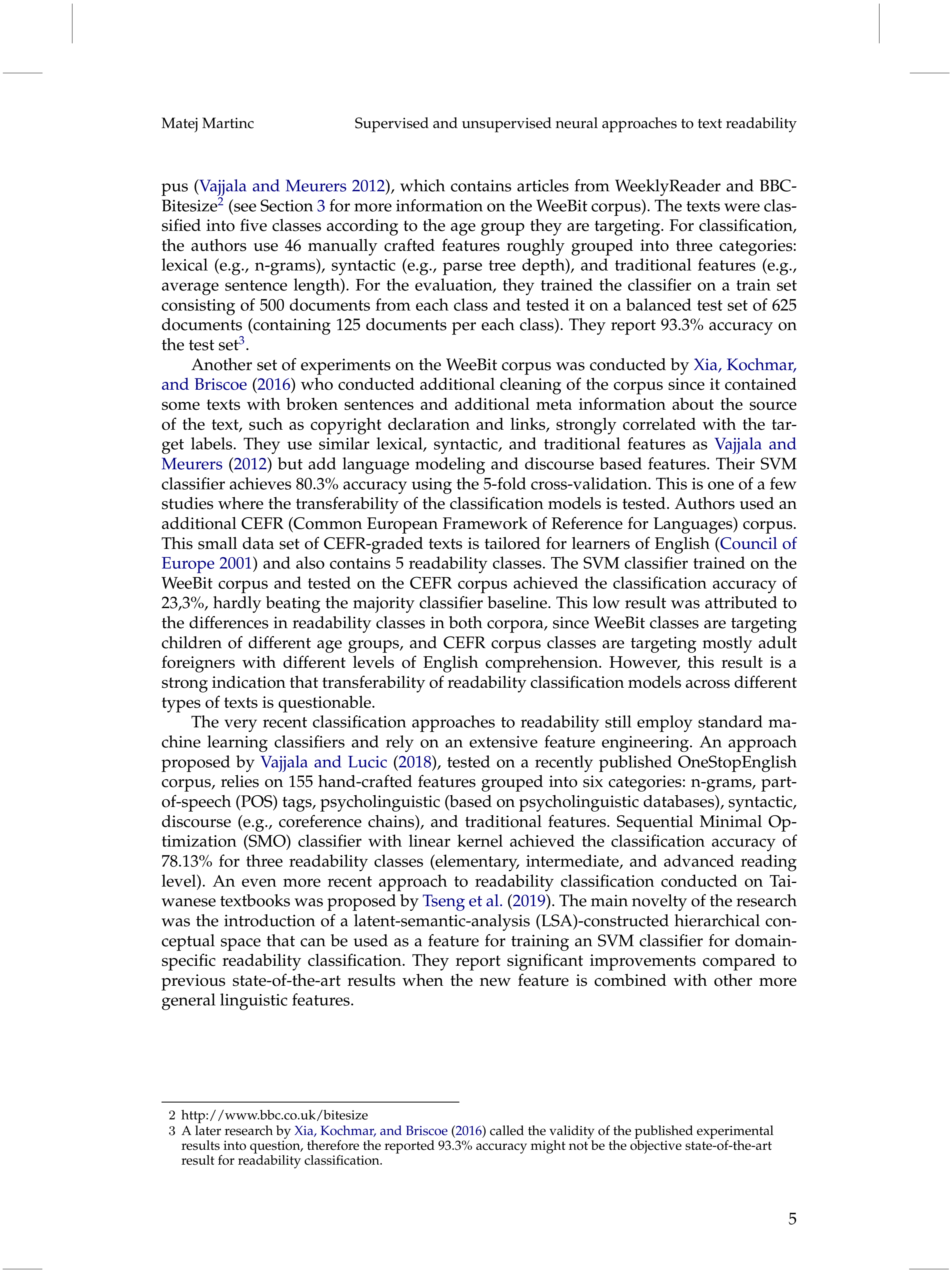
2.2 Classification approach to readability

The alternative to measuring readability with statistical formulas is to consider it a prediction task and predict the level of readability. These approaches usually require extensive feature engineering and thereby address some deficiencies of statistical formulas, such as their reductionism and dismissal of contextual and semantic information.

One of the first classification approaches to readability was proposed by Schwarm and Ostendorf (2005). It relies on a Support Vector Machine (SVM) classifier trained on a WeeklyReader corpus[1], containing articles grouped into four classes according to the age of the target audience. Statistical language models, statistical readability formulas, and parse trees are used as features in the model. This approach was extended and improved upon in Petersen and Ostendorf (2009).

A successful classification approach to readability was proposed by Vajjala and Meurers (2012). Their multi-layer perceptron classifier is trained on the WeeBit cor-

[1] http://www.weeklyreader.com



pus (Vajjala and Meurers 2012), which contains articles from WeeklyReader and BBC-Bitesize[2] (see Section 3 for more information on the WeeBit corpus). The texts were classified into five classes according to the age group they are targeting. For classification, the authors use 46 manually crafted features roughly grouped into three categories: lexical (e.g., n-grams), syntactic (e.g., parse tree depth), and traditional features (e.g., average sentence length). For the evaluation, they trained the classifier on a train set consisting of 500 documents from each class and tested it on a balanced test set of 625 documents (containing 125 documents per each class). They report 93.3% accuracy on the test set[3].

Another set of experiments on the WeeBit corpus was conducted by Xia, Kochmar, and Briscoe (2016) who conducted additional cleaning of the corpus since it contained some texts with broken sentences and additional meta information about the source of the text, such as copyright declaration and links, strongly correlated with the target labels. They use similar lexical, syntactic, and traditional features as Vajjala and Meurers (2012) but add language modeling and discourse based features. Their SVM classifier achieves 80.3% accuracy using the 5-fold cross-validation. This is one of a few studies where the transferability of the classification models is tested. Authors used an additional CEFR (Common European Framework of Reference for Languages) corpus. This small data set of CEFR-graded texts is tailored for learners of English (Council of Europe 2001) and also contains 5 readability classes. The SVM classifier trained on the WeeBit corpus and tested on the CEFR corpus achieved the classification accuracy of 23,3%, hardly beating the majority classifier baseline. This low result was attributed to the differences in readability classes in both corpora, since WeeBit classes are targeting children of different age groups, and CEFR corpus classes are targeting mostly adult foreigners with different levels of English comprehension. However, this result is a strong indication that transferability of readability classification models across different types of texts is questionable.

The very recent classification approaches to readability still employ standard machine learning classifiers and rely on an extensive feature engineering. An approach proposed by Vajjala and Lucic (2018), tested on a recently published OneStopEnglish corpus, relies on 155 hand-crafted features grouped into six categories: n-grams, part-of-speech (POS) tags, psycholinguistic (based on psycholinguistic databases), syntactic, discourse (e.g., coreference chains), and traditional features. Sequential Minimal Optimization (SMO) classifier with linear kernel achieved the classification accuracy of 78.13% for three readability classes (elementary, intermediate, and advanced reading level). An even more recent approach to readability classification conducted on Taiwanese textbooks was proposed by Tseng et al. (2019). The main novelty of the research was the introduction of a latent-semantic-analysis (LSA)-constructed hierarchical conceptual space that can be used as a feature for training an SVM classifier for domain-specific readability classification. They report significant improvements compared to previous state-of-the-art results when the new feature is combined with other more general linguistic features.

[2] http://www.bbc.co.uk/bitesize

[3] A later research by Xia, Kochmar, and Briscoe (2016) called the validity of the published experimental results into question, therefore the reported 93.3% accuracy might not be the objective state-of-the-art result for readability classification.

2.3 Statistical language models

The standard task of language modeling can be formally defined as predicting a probability distribution of words from the fixed size vocabulary V, for word $w_{\text{t+1}}$, given the historical sequence $w_{\text{1:t}} = [w_1, ..., w_t]$. From a statistical point of view, taking an entire historical sequence of words into consideration is problematic due to data sparsity, since the majority of possible word sequences will not be observed in the training sample. In order to handle sequences that were not seen during training, the standard solution (called the n-gram language model) limits the historical sequence to n previous words, counts the observed n-grams, and employs any of a number of different smoothing techniques (Chen and Goodman 1999). A special version of the n-gram model is a unigram model ($n = 1$), where the probability of each word depends only on that word's probability in the document. A recent solution to data sparsity is the introduction of neural language models (Mikolov et al. 2011), which will be explained in Section 2.4.

To measure the performance of language models, traditionally a metric called perplexity is used. A language model m is evaluated according to how well it predicts a separate test sequence of words $w_{\text{1:N}} = [w_1, ..., w_N]$. For this case, the perplexity (PPL) of the language model $m()$ is defined as:

$$\text{PPL} = 2^{-\frac{1}{N}\sum_{i=1}^{N}\log_2 m(w_i)}, \tag{1}$$

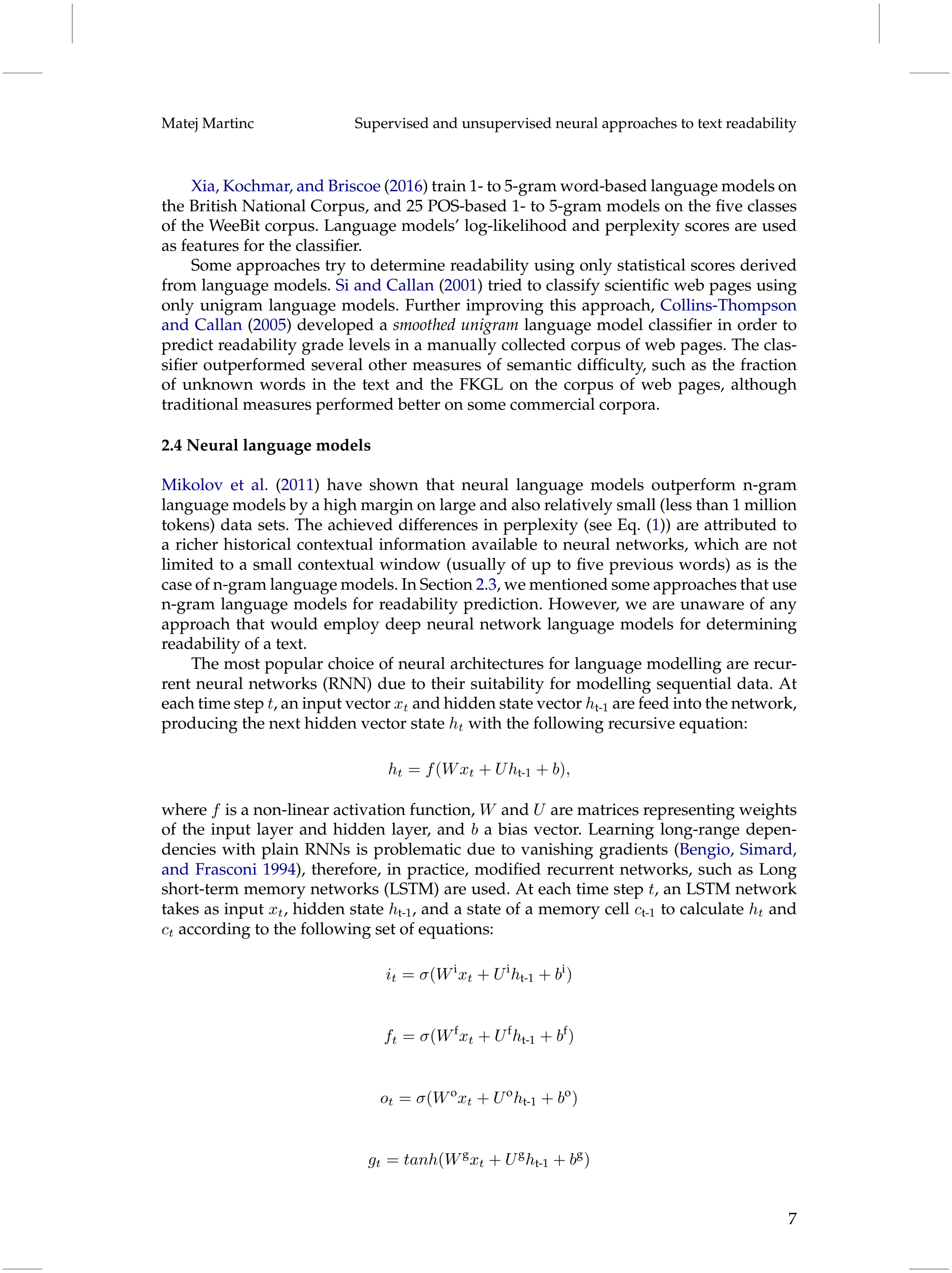
where $m(w_i)$ is the probability assigned to word w_i by the language model m, and N is the length of the sequence. The lower the perplexity score, the better the language model predicts the words in a document, i.e. the more predictable and aligned with the training set the text is.

Of special interest to our method are features of language models, used by many classification approaches (see Section 2.2 above). Schwarm and Ostendorf (2005) train one n-gram language model for each readability class *c* in the training data set. For each text document *d*, they calculate the likelihood ratio according to the following formula:

$$LR(d, c) = \frac{P(d|c)P(c)}{\sum_{\bar{c} \neq c} P(d|\bar{c})P(\bar{c})},$$

where $P(d|c)$ denotes the probability returned by the language model trained on texts labeled with class *c*, and $P(d|\bar{c})$ denotes probability of *d* returned by the language model trained on the class $\bar{c}$. Uniform prior probabilities of classes are assumed. The likelihood ratios are used as features in the classification model along with perplexities achieved by all the models.

In Petersen and Ostendorf (2009), three statistical language models (unigram, bigram and trigram) are trained on four external data resources: Britannica (adult), Britannica Elementary, CNN (adult) and CNN abridged. The resulting twelve n-gram language models are used to calculate perplexities of each target document. It is assumed that low perplexity scores calculated by the language models trained on the adult level texts and high perplexity scores calculated by the language models trained on the elementary/abridged levels would indicate a high reading level, and high perplexity scores calculated by the language models trained on the adult level texts and low perplexity scores calculated by the language models trained on the elementary/abridged levels would indicate a low reading level.



Xia, Kochmar, and Briscoe (2016) train 1- to 5-gram word-based language models on the British National Corpus, and 25 POS-based 1- to 5-gram models on the five classes of the WeeBit corpus. Language models' log-likelihood and perplexity scores are used as features for the classifier.

Some approaches try to determine readability using only statistical scores derived from language models. Si and Callan (2001) tried to classify scientific web pages using only unigram language models. Further improving this approach, Collins-Thompson and Callan (2005) developed a *smoothed unigram* language model classifier in order to predict readability grade levels in a manually collected corpus of web pages. The classifier outperformed several other measures of semantic difficulty, such as the fraction of unknown words in the text and the FKGL on the corpus of web pages, although traditional measures performed better on some commercial corpora.

2.4 Neural language models

Mikolov et al. (2011) have shown that neural language models outperform n-gram language models by a high margin on large and also relatively small (less than 1 million tokens) data sets. The achieved differences in perplexity (see Eq. (1)) are attributed to a richer historical contextual information available to neural networks, which are not limited to a small contextual window (usually of up to five previous words) as is the case of n-gram language models. In Section 2.3, we mentioned some approaches that use n-gram language models for readability prediction. However, we are unaware of any approach that would employ deep neural network language models for determining readability of a text.

The most popular choice of neural architectures for language modelling are recurrent neural networks (RNN) due to their suitability for modelling sequential data. At each time step t, an input vector x_t and hidden state vector $h_{\text{t-1}}$ are feed into the network, producing the next hidden vector state h_t with the following recursive equation:

$$h_t = f(Wx_t + Uh_{\text{t-1}} + b),$$

where f is a non-linear activation function, W and U are matrices representing weights of the input layer and hidden layer, and b a bias vector. Learning long-range dependencies with plain RNNs is problematic due to vanishing gradients (Bengio, Simard, and Frasconi 1994), therefore, in practice, modified recurrent networks, such as Long short-term memory networks (LSTM) are used. At each time step t, an LSTM network takes as input x_t, hidden state $h_{\text{t-1}}$, and a state of a memory cell $c_{\text{t-1}}$ to calculate h_t and c_t according to the following set of equations:

$$i_t = \sigma(W^{\mathrm{i}}x_t + U^{\mathrm{i}}h_{\text{t-1}} + b^{\mathrm{i}})$$

$$f_t = \sigma(W^{\mathrm{f}}x_t + U^{\mathrm{f}}h_{\text{t-1}} + b^{\mathrm{f}})$$

$$o_t = \sigma(W^{\mathrm{o}}x_t + U^{\mathrm{o}}h_{\text{t-1}} + b^{\mathrm{o}})$$

$$g_t = tanh(W^{\mathrm{g}}x_t + U^{\mathrm{g}}h_{\text{t-1}} + b^{\mathrm{g}})$$

$$c_t = f_t \odot c_{\text{t-1}} + i_t \odot g_t$$

$$h_t = o_t \odot tanh(c_t),$$

where i_t, f_t and o_t are reffered to as input, forget and output gates, respectively. σ and $tanh$ are element-wise sigmoid and hyperbolic tangent functions and $\odot$ represents a dot product operation.

In our experiments, we use the LSTM-based language model proposed by Kim et al. (2016). This system is adapted to language modelling of morphologically rich languages, such as Slovenian, by employing an additional character level convolutional neural network (CNN). The convolutional level learns a character structure of words and is connected to the LSTM-based language model, which produces predictions at the word level.

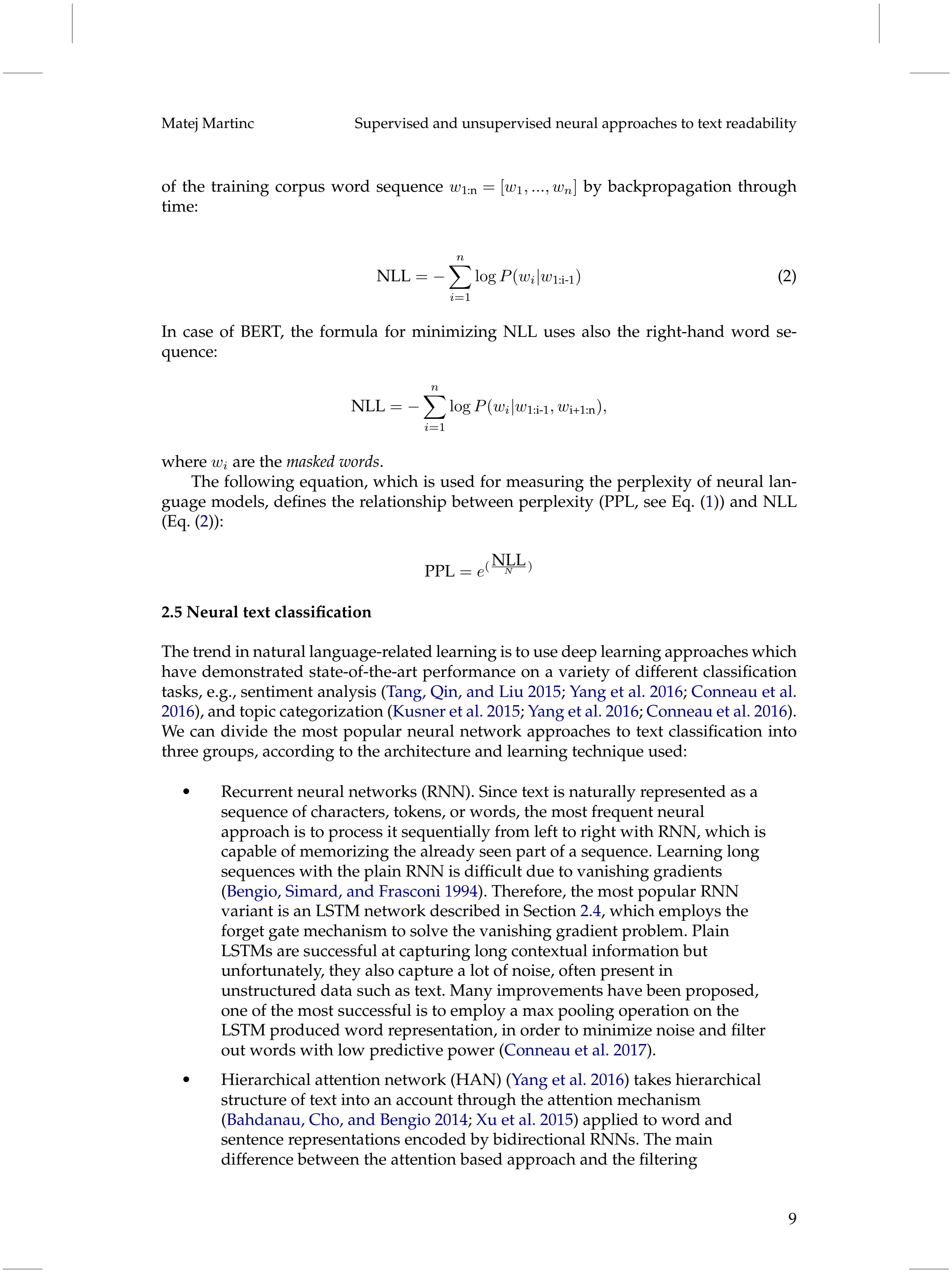
Recently, Bai, Kolter, and Koltun (2018) introduced a new sequence modelling architecture based on convolution, called temporal convolutional network (TCN). TCN uses casual convolution operations, which make sure that there is no information leakage from future time steps to the past. This and the fact that TCN takes a sequence as an input and maps it into an output sequence of the same size, makes this architecture appropriate for language modelling. TCNs are capable of leveraging long contexts for their prediction by using a very deep network architecture and a hierarchy of dilated convolutions. A single dilated convolution operation F on element s of the 1-dimensional sequence x can be defined with the following equation:

$$F(s) = (x *_{\text{d}} f)(s) = \sum_{i=0}^{k-1} f(i) \cdot x_{s-d\cdot i},$$

where $f : 0, \dots k-1$ is a filter of size k, d a dilation factor and $s - d \cdot i$ accounts for the direction of the past. In this way, the context taken into account during the prediction can be increased by using larger filter sizes and by increasing the dilation factor. The most common practice is to increase the dilation factor exponentially with the depth of the network.

Another recent approach to language modelling was proposed by Devlin et al. (2018). The BERT (Bidirectional Encoder Representations from Transformers) uses both left and right context, which means that a word w_{t} in a sequence is not determined just from its left sequence $w_{\text{1:t-1}} = [w_1, ..., w_{t-1}]$ but also from its right word sequence $w_{\text{t+1:n}} = [w_{t+1}, ..., w_{t+n}]$. This approach introduces a new learning objective, a *masked language model*, where a predefined percentage of randomly chosen words from the input word sequence are masked, and the objective is to predict these masked words from the unmasked context. This approach uses a transformer architecture, which relies on a self-attention mechanism proposed by Vaswani et al. (2017). The distinguishing feature of this approach is the employment of several parallel attention layers, the so-called *attention heads*, which reduce the computational cost and allow the system to attend to several dependencies at once.

All types of neural network language models, TCN, LSTM, and BERT, output softmax probability distribution calculated over the entire vocabulary, and present the probabilities for each word given its historical (and in case of BERT also future) sequence. Training of these networks usually minimizes the negative log-likelihood (NLL)



of the training corpus word sequence $w_{1:\mathrm{n}} = [w_1, ..., w_n]$ by backpropagation through time:

$$\mathrm{NLL} = -\sum_{i=1}^{n} \log P(w_i | w_{\mathrm{1:i\text{-}1}}) \tag{2}$$

In case of BERT, the formula for minimizing NLL uses also the right-hand word sequence:

$$\mathrm{NLL} = -\sum_{i=1}^{n} \log P(w_i | w_{\mathrm{1:i\text{-}1}}, w_{\mathrm{i+1:n}}),$$

where w_i are the *masked words*.

The following equation, which is used for measuring the perplexity of neural language models, defines the relationship between perplexity (PPL, see Eq. (1)) and NLL (Eq. (2)):

$$\mathrm{PPL} = e^{(\frac{\mathrm{NLL}}{N})}$$

2.5 Neural text classification

The trend in natural language-related learning is to use deep learning approaches which have demonstrated state-of-the-art performance on a variety of different classification tasks, e.g., sentiment analysis (Tang, Qin, and Liu 2015; Yang et al. 2016; Conneau et al. 2016), and topic categorization (Kusner et al. 2015; Yang et al. 2016; Conneau et al. 2016). We can divide the most popular neural network approaches to text classification into three groups, according to the architecture and learning technique used:

- Recurrent neural networks (RNN). Since text is naturally represented as a sequence of characters, tokens, or words, the most frequent neural approach is to process it sequentially from left to right with RNN, which is capable of memorizing the already seen part of a sequence. Learning long sequences with the plain RNN is difficult due to vanishing gradients (Bengio, Simard, and Frasconi 1994). Therefore, the most popular RNN variant is an LSTM network described in Section 2.4, which employs the forget gate mechanism to solve the vanishing gradient problem. Plain LSTMs are successful at capturing long contextual information but unfortunately, they also capture a lot of noise, often present in unstructured data such as text. Many improvements have been proposed, one of the most successful is to employ a max pooling operation on the LSTM produced word representation, in order to minimize noise and filter out words with low predictive power (Conneau et al. 2017).
- Hierarchical attention network (HAN) (Yang et al. 2016) takes hierarchical structure of text into an account through the attention mechanism (Bahdanau, Cho, and Bengio 2014; Xu et al. 2015) applied to word and sentence representations encoded by bidirectional RNNs. The main difference between the attention based approach and the filtering

approach proposed by Conneau et al. (2017), is the acknowledgment, that the informativeness of words and sentences is context-dependent, therefore the same words and sentences in different documents might have a completely different predictive power.
Given a sentence s_i with the word representation $u_{it}, t \in [0, T]$, the attention mechanism on the word level can be described with the following set of equations:

$$u_{it} = tanh(W_w h_{it} + b_w),$$

$$\alpha_{it} = \frac{exp(u_{it}^T u_c)}{\sum_t exp(u_{it}^T u_c)},$$

$$s_i = \sum_i \alpha_{it} h_{it}.$$

The word representation h_{it} is first fed to a dense layer W_w with the $tanh$ activation function to get a hidden representation u_{it}. The importance of the hidden representation is calculated by measuring the similarity between the u_{it} and randomly initialized context vector u_c. The softmax function is applied to derive a normalized similarity weight α_{it}, which is used for calculation of the final sequence vector s_i as a weighted sum of the h_{it}. The final sequence vector s_i, calculated on the word level, is used as an input to the same attention mechanism on the sentence level, which produces a document representation as an output. This output is used as a feature matrix for the final document classification.

- Transfer learning is the latest state-of-the-art approach to text classification (Howard and Ruder 2018; Devlin et al. 2018). In this approach, we first pretrain a neural language model on a large general corpora and then fine-tune this model for a specific classification task by adding the final classification layer. The network with an additional layer is trained for a few additional epochs on new data. The syntactic and semantic knowledge of the pretrained language model is transferred and leveraged for the new classification task. An example of this approach is the BERT language model (Devlin et al. 2018) pretrained on the concatenation of BooksCorpus (800M words) (Zhu et al. 2015) and English Wikipedia (2,500M words), to which an additional linear classification head is added. This model achieved state-of-the-art results on many text classification tasks, such as the question answering task on the SQuAD dataset (Rajpurkar et al. 2016), and several language inference tasks.

3. Datasets

All experiments are conducted on four corpora labelled with readability scores:

- **The WeeBit corpus**: The articles from WeeklyReader[4] and BBC-Bitesize[5] are classified into five classes according to the age group they are targeting. The classes correspond to age groups between 7-8, 8-9, 9-10, 10-14 and 14-16. In the original corpus of Vajjala and Meurers (2012), the classes are balanced and the corpus contains altogether 3125 documents, 625 per class. In our experiments, we followed recommendations of Xia, Kochmar, and Briscoe (2016) in order to fix broken sentences and remove additional meta information, such as copyright declaration and links, strongly correlated with the target labels. We reextracted the corpus from the HTML files according to the procedure described in Xia, Kochmar, and Briscoe (2016) and discarded some documents due to the lack of content after the extraction and cleaning process. The final corpus used in our experiments contains altogether 3000 documents, 600 per class.
- **The OneStopEnglish corpus** (Vajjala and Lucic 2018) contains aligned texts of three distinct reading levels (beginner, intermediate, and advanced) that were written specifically for English as Second Language (ESL) learners. The corpus consists of 189 texts, each written in three versions (567 in total). The corpus is freely available[6].
- **The Newsela corpus** (Xu, Callison-Burch, and Napoles 2015). We use the version of the corpus from 29 January 2016 consisting of altogether 10,786 documents, out of which we only used 9,565 English documents. The corpus contains 1,911 original English articles and up to five simplified versions for every original article. The original and simplified versions correspond to altogether eleven different grade levels (from 2nd to 12th grade). Grade levels are imbalanced; the exact numbers of articles per grade are presented in Table 1.
- **Corpus of Slovenian school books (Slovenian SB):** In order to test the transferability of the proposed approaches to other languages, a corpus of Slovenian school books was compiled. The corpus contains 3,639,665 words in 125 school books for nine grades of primary schools and four grades of secondary school. For supervised classification experiments, we split the school books into chunks twenty sentences long, in order to build a train and test set with sufficient number of documents. The exact number of school books and chunks per grade are presented in Table 2.

Language models are trained on large corpora of texts. We used the following corpora.

- **Corpus of English Wikipedia** and **Corpus of Simple Wikipedia** articles. We created three corpora for the use in our unsupervised English experiments[7]:

[4] http://www.weaklyreader.com
[5] http://www.bbc.co.uk/bitesize
[6] https://zenodo.org/record/1219041
[7] English Wikipedia and Simple Wikipedia dumps from 26th of January 2018 were used for the corpus construction

Table 1
The number of English articles and tokens per specific grade in the Newsela corpus.

Grade	#documents	#tokens
2nd	224	74,428
3rd	500	197,992
4th	1,569	923,828
5th	1,342	912,411
6th	1,058	802,057
7th	1,210	979,471
8th	1,037	890,358
9th	750	637,784
10th	20	19,012
11th	2	1,130
12th	1,853	1,833,781
all	9,565	7,272,252

  - **Wiki-normal** contains 130,000 randomly selected articles from the Wikipedia dump;
  - **Wiki-simple** contains 130,000 randomly selected articles from the Simple Wikipedia dump;
  - **Wiki-balanced** contains 65,000 randomly selected articles from the Wikipedia dump (dated 26 January 2018) and 65,000 randomly selected articles from the Simple Wikipedia dump.

- **KRES-balanced:** KRES corpus (Logar et al. 2012) is a 100 million word balanced reference corpus of Slovenian language. 35% of its content are books, 40% periodicals, and 20% internet texts. From this corpus we took all the available documents from two children magazines (Ciciban and Cicido), all documents from four teenager magazines (Cool, Frka, PIL plus and Smrklja), and documents from three magazines targeting adult audiences (Življenje in tehnika, Radar, City magazine). With these texts we built a corpus with approximately 2.4 million words. The corpus is balanced in a sense that about one third of the sentences come from documents targeting children, one third is targeting teenagers, and the last third is targeting adults.

4. Unsupervised neural approach

In this section, we explore how language models can be used for determining readability of the text by injecting discourse cohesion and background knowledge information into the measurement of readability. In Section 4.1, we describe the methodology, and in Section 4.2.2, we present the results of the conducted experiments.

4.1 Methodology

The main tool we use for assessment of readability in an unsupervised setting are neural language models, described in Section 2.4. We use three types of architectures for neural language models, recurrent (LSTM), convolutional, and transformer neural

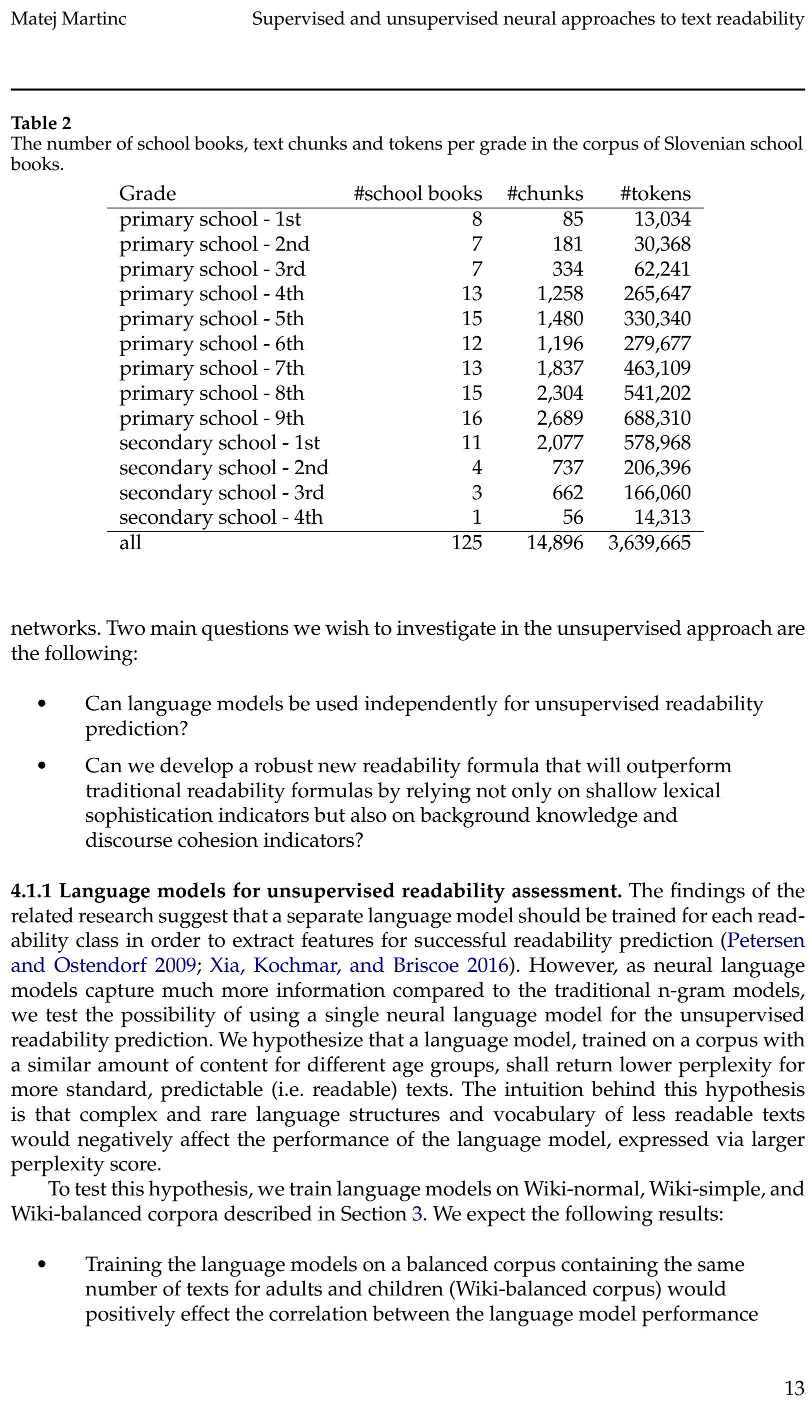


Table 2
The number of school books, text chunks and tokens per grade in the corpus of Slovenian school books.

Grade	#school books	#chunks	#tokens
primary school - 1st	8	85	13,034
primary school - 2nd	7	181	30,368
primary school - 3rd	7	334	62,241
primary school - 4th	13	1,258	265,647
primary school - 5th	15	1,480	330,340
primary school - 6th	12	1,196	279,677
primary school - 7th	13	1,837	463,109
primary school - 8th	15	2,304	541,202
primary school - 9th	16	2,689	688,310
secondary school - 1st	11	2,077	578,968
secondary school - 2nd	4	737	206,396
secondary school - 3rd	3	662	166,060
secondary school - 4th	1	56	14,313
all	125	14,896	3,639,665

networks. Two main questions we wish to investigate in the unsupervised approach are the following:

- Can language models be used independently for unsupervised readability prediction?
- Can we develop a robust new readability formula that will outperform traditional readability formulas by relying not only on shallow lexical sophistication indicators but also on background knowledge and discourse cohesion indicators?

4.1.1 Language models for unsupervised readability assessment. The findings of the related research suggest that a separate language model should be trained for each readability class in order to extract features for successful readability prediction (Petersen and Ostendorf 2009; Xia, Kochmar, and Briscoe 2016). However, as neural language models capture much more information compared to the traditional n-gram models, we test the possibility of using a single neural language model for the unsupervised readability prediction. We hypothesize that a language model, trained on a corpus with a similar amount of content for different age groups, shall return lower perplexity for more standard, predictable (i.e. readable) texts. The intuition behind this hypothesis is that complex and rare language structures and vocabulary of less readable texts would negatively affect the performance of the language model, expressed via larger perplexity score.

To test this hypothesis, we train language models on Wiki-normal, Wiki-simple, and Wiki-balanced corpora described in Section 3. We expect the following results:

- Training the language models on a balanced corpus containing the same number of texts for adults and children (Wiki-balanced corpus) would positively effect the correlation between the language model performance and readability, since all our test corpora (WeeBit, OneStopEnglish and Newsela) contain texts meant for children and young adults.

- The language models trained only on texts for adults (Wiki-normal) will show higher perplexity on texts for children, since their training set did not contain such texts; this will negatively effect the correlation between the language model performance and readability.
- Training the language models only on texts for children (Wiki-simple corpus) will result in a higher perplexity score of the language model when applied to adult texts. This will positively effect the correlation between the language models' performance and readability. However, this language model will not be able to reliably distinguish between texts for different age groups of young adults and teenagers, which will have a negative effect on the correlation.

Note that all three Wiki corpora contain the same amount of articles, in order to make sure that the training set size does not influence the results of the experiments.

To further test the viability of the hypothesis presented above and to test the limits of using a single language model for unsupervised readability prediction, we also explore the possibility of using a language model trained on a large general corpus for the unsupervised readability prediction.

4.1.2 Ranked sentence readability score. Based on the two considerations below, we propose a new Ranked Sentence Readability Score (RSRS) for measuring the readability with language models.

- The shallow lexical sophistication indicators, such as the length of a sentence, correlate well with the readability of a text. Using them besides statistics derived from language models could improve the unsupervised readability prediction.
- The perplexity score used for measuring the performance of a language model is an *unweighted* sum of perplexities of words in the predicted sequence. In reality, a small amount of unreadable words might drastically reduce the readability of the entire text. Assigning larger weights to such words might improve the correlation of language model scores with the readability.

The proposed readability score is calculated with the following procedure. First, a given text is split into sentences with the default sentence tokenizer from the NLTK library (Bird and Loper 2004). In order to get a readability estimation for each word in a specific context, we compute, for each word in the sentence, the word negative log-likelihood (WNLL) according to the following formula:

$$\text{WNLL} = -(y_t \log y_p + (1 - y_t) \log (1 - y_p)),$$

where y_p denotes the probability (from the softmax distribution) predicted by the language model according to the historical sequence, and y_t denotes the true probability distribution of a word. The y_t has the value 1 for the word in the vocabulary that actually appears next in the sequence and the value 0 for all the other words in the vocabulary.

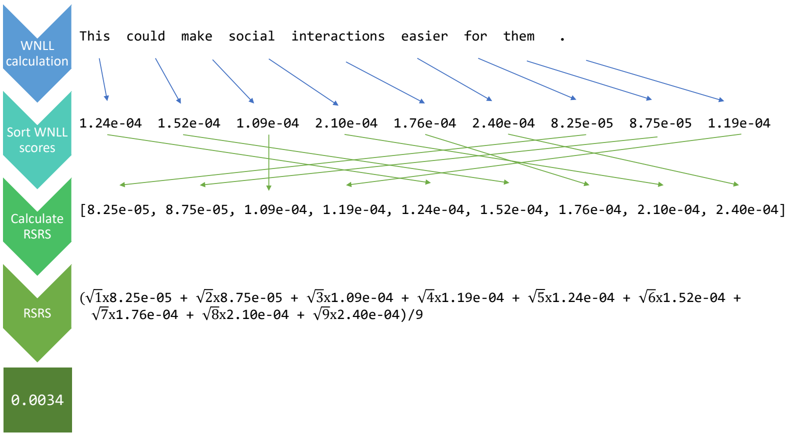

WNLL calculation
This could make social interactions easier for them .
Sort WNLL scores
1.24e-04 1.52e-04 1.09e-04 2.10e-04 1.76e-04 2.40e-04 8.25e-05 8.75e-05 1.19e-04
Calculate RSRS
[8.25e-05, 8.75e-05, 1.09e-04, 1.19e-04, 1.24e-04, 1.52e-04, 1.76e-04, 2.10e-04, 2.40e-04]
RSRS
(√1x8.25e-05 + √2x8.75e-05 + √3x1.09e-04 + √4x1.19e-04 + √5x1.24e-04 + √6x1.52e-04 + √7x1.76e-04 + √8x2.10e-04 + √9x2.40e-04)/9
0.0034

Figure 1
The RSRS calculation for the sentence *This could make social interactions easier for them.*

Next, we sort all the words in the sentence in ascending order according to their WNLL score and the ranked sentence readability score (RSRS) is calculated with the following expression:

$$\mathrm{RSRS} = \frac{\sum_{i=1}^{S} \sqrt{i} \cdot \mathrm{WNLL}(i)}{S}, \tag{3}$$

where S denotes the sentence length and i represents the rank of a word in a sentence according to its WNLL value. The square root of the word rank is used for proportionally weighting words according to their readability, since initial experiments suggested that the use of a square root of a rank represents the best balance between allowing all words to contribute equally to the overall readability of the sentence and allowing only the least readable words to affect the overall readability of the sentence. For out of vocabulary words, square root rank weights are doubled, since these rare words are in our opinion good indicators of non-standard text. Finally, in order to get the readability score for the entire text, we calculate the average of all the RSRS scores in the text. An example of how RSRS is calculated for a specific sentence is shown in Figure 1.

The main idea behind the RSRS score is to avoid the reductionism of traditional readability formulas. We aim to achieve this by including discourse cohesion and background knowledge through language model based statistics. The first assumption is that low discourse cohesion has a negative effect on the performance of the language model, resulting in a higher WNLL for words in complex grammatical and lexical contexts. The second assumption is that the background knowledge is included in the readability calculation: tested documents with semantics dissimilar to the documents in the language model training set will negatively affect the performance of the language model, resulting in the higher WNLL score for words with unknown semantics. The trainable nature of language models allows for customization and personalization of the RSRS for specific tasks, topics and languages. This means that RSRS shall alleviate the problem of cultural non-transferability of traditional readability formulas.

On the other hand, the RSRS also leverages shallow lexical sophistication indicators through the index weighting scheme which makes sure that less readable words contribute more to the overall readability score. This is somewhat similar to the counts of long and difficult words in the traditional readability formulas, such as GFI and DCRF. The value of RSRS also increases for texts containing longer sentences, since the square roots of the word rank weights become larger with increased sentence length. This is similar to the behaviour of traditional formulas such as GFI, FRE, FKGL, ARI, DCRF, where this effect is achieved by incorporating the ratio between the total number of words and the total number of sentences into the equation.

4.2 Unsupervised experiments

For the presented unsupervised readability assessment methodology based on neural language models, we first present the experimental design followed by the results.

4.2.1 Experimental design. Three different architectures of language models (described in Section 2.4) are used for experiments: a convolutional word level language model (CLM) proposed by Bai, Kolter, and Koltun (2018), a recurrent language model (RLM) proposed by Kim et al. (2016), and an attention based language model BERT (Devlin et al. 2018). For the experiments on English language, we train CLM and RLM on three Wiki corpora. To explore the possibility of using a language model trained on a general corpus for the unsupervised readability prediction, we use a pretrained BERT language model trained on the Google Books Corpus (Goldberg and Orwant 2013) (800M words) and Wikipedia (2,500M words) for the experiments on English. For the experiments on Slovenian language, corpora containing just texts for children are too small for efficient training of language models, therefore CLM and RLM were only trained on the KRES-balanced corpus described in Section 3. For exploring the possibility of using a general language model for the unsupervised readability prediction, a pretrained BERT multilingual language model trained on Wikipedia dumps of hundred languages with the largest Wikipedia, including Slovenian, is used.

The performance of language models is typically measured with the perplexity (see Eq. (1)). To answer the research question if language models can be used independently for unsupervised readability prediction, we investigate how the measured perplexity of language models correlates with the readability labels in the gold-standard WeeBit, OneStopEnglish, Newsela, and Slovenian school books corpora described in Section 3. The correlation to these ground truth readability labels is also used to evaluate the performance of the RSRS measure. For performance comparison, we calculate the traditional readability formula values (described in Section 2) for each document in the gold-standard corpora and also measure the correlation between these values and manually assigned labels. As a baseline we use the average sentence length in each document.

The correlation is measured with the Pearson correlation coefficient (ρ). Given a pair of distributions X and Y, the covariance cov, and the standard deviation σ, the formula for ρ is:

$$\rho_{x,y} = \frac{cov(x, y)}{\sigma_x \sigma_y}$$

A larger positive correlation signifies a better performance for all measures except the FRE readability measure. As this formula assigns higher scores to better readable texts, a larger negative correlation suggests a better performance of the measure.

4.2.2 Experimental results. The results of the experiments are presented in Table 3. The average ranking of measures on English and Slovenian datasets are presented in Table 4.

The correlation coefficient for all measures vary drastically between different corpora. The highest ρ values are obtained on the Newsela corpus, where the best performing measure (surprisingly this is our baseline - the average sentence length) achieves the ρ of 0.906. The highest ρ on the other two English corpora are much lower. On the WeeBit corpus, the best performance is achieved by GFI and FKGL measures (ρ of 0.544) and on the OneStopEnglish corpus the best performance is achieved with the proposed CLM RSRS-simple (ρ of 0.615). On the Slovenian school books, the ρ values are higher and the best performing measure is CLM RSRS score-balanced with ρ of 0.789.

The perplexity-based measures show much lower correlation with the ground truth readability scores. Overall, they perform the worst of all the measures for both languages (see Table 4) but we can observe large differences in their performance across different corpora. While there is either no correlation or low negative correlation between perplexities of all three language models and readability on the WeeBit corpus, there is

Table 3
Pearson correlation coefficient between manually assigned readability labels and the readability scores assigned by different readability measures in the unsupervised setting. The highest correlation for each corpus is marked with the bold typeface.

Measure/Dataset	WeeBit	OneStopEnglish	Newsela	Slovenian SB
RLM perplexity-balanced	-0.0819	0.405	0.512	0.303
RLM perplexity-simple	-0.115	0.420	0.470	/
RLM perplexity-normal	-0.127	0.283	0.341	/
CLM perplexity-balanced	-0.0402	0.474	0.528	0.136
CLM perplexity-simple	-0.0542	0.524	0.583	/
CLM perplexity-normal	-0.117	0.292	0.270	/
BERT perplexity	-0.123	-0.162	-0.673	-0.651
RLM RSRS-balanced	0.497	0.551	0.890	0.732
RLM RSRS-simple	0.506	0.569	0.893	/
RLM RSRS-normal	0.490	0.536	0.886	/
CLM RSRS-balanced	0.446	0.599	0.894	**0.789**
CLM RSRS-simple	0.451	**0.615**	0.896	/
CLM RSRS-normal	0.414	0.576	0.890	/
BERT RSRS	0.279	0.384	0.674	-0.301
GFI	**0.544**	0.550	0.849	0.730
FRE	-0.433	-0.485	-0.775	-0.614
FKGL	**0.544**	0.533	0.865	0.697
ARI	0.488	0.520	0.875	0.658
DCRF	0.420	0.496	0.735	0.686
SMOG	0.456	0.498	0.813	0.770
Avg. sentence length	0.508	0.498	**0.906**	0.683

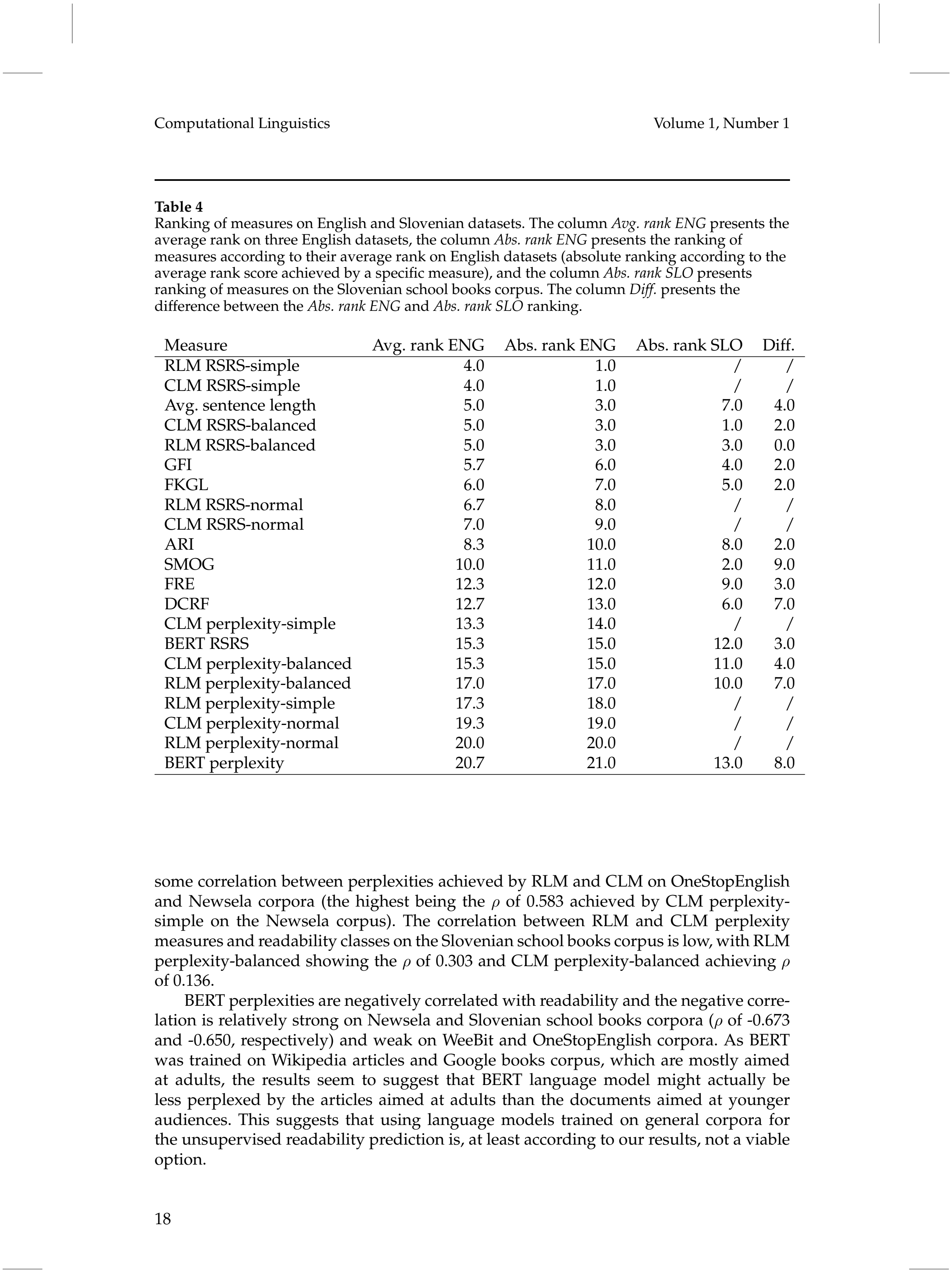


Table 4
Ranking of measures on English and Slovenian datasets. The column *Avg. rank ENG* presents the average rank on three English datasets, the column *Abs. rank ENG* presents the ranking of measures according to their average rank on English datasets (absolute ranking according to the average rank score achieved by a specific measure), and the column *Abs. rank SLO* presents ranking of measures on the Slovenian school books corpus. The column *Diff.* presents the difference between the *Abs. rank ENG* and *Abs. rank SLO* ranking.

Measure	Avg. rank ENG	Abs. rank ENG	Abs. rank SLO	Diff.
RLM RSRS-simple	4.0	1.0	/	/
CLM RSRS-simple	4.0	1.0	/	/
Avg. sentence length	5.0	3.0	7.0	4.0
CLM RSRS-balanced	5.0	3.0	1.0	2.0
RLM RSRS-balanced	5.0	3.0	3.0	0.0
GFI	5.7	6.0	4.0	2.0
FKGL	6.0	7.0	5.0	2.0
RLM RSRS-normal	6.7	8.0	/	/
CLM RSRS-normal	7.0	9.0	/	/
ARI	8.3	10.0	8.0	2.0
SMOG	10.0	11.0	2.0	9.0
FRE	12.3	12.0	9.0	3.0
DCRF	12.7	13.0	6.0	7.0
CLM perplexity-simple	13.3	14.0	/	/
BERT RSRS	15.3	15.0	12.0	3.0
CLM perplexity-balanced	15.3	15.0	11.0	4.0
RLM perplexity-balanced	17.0	17.0	10.0	7.0
RLM perplexity-simple	17.3	18.0	/	/
CLM perplexity-normal	19.3	19.0	/	/
RLM perplexity-normal	20.0	20.0	/	/
BERT perplexity	20.7	21.0	13.0	8.0

some correlation between perplexities achieved by RLM and CLM on OneStopEnglish and Newsela corpora (the highest being the ρ of 0.583 achieved by CLM perplexity-simple on the Newsela corpus). The correlation between RLM and CLM perplexity measures and readability classes on the Slovenian school books corpus is low, with RLM perplexity-balanced showing the ρ of 0.303 and CLM perplexity-balanced achieving ρ of 0.136.

BERT perplexities are negatively correlated with readability and the negative correlation is relatively strong on Newsela and Slovenian school books corpora (ρ of -0.673 and -0.650, respectively) and weak on WeeBit and OneStopEnglish corpora. As BERT was trained on Wikipedia articles and Google books corpus, which are mostly aimed at adults, the results seem to suggest that BERT language model might actually be less perplexed by the articles aimed at adults than the documents aimed at younger audiences. This suggests that using language models trained on general corpora for the unsupervised readability prediction is, at least according to our results, not a viable option.

In regards to our hypothesis that a language model trained on a corpus with similar amount of content for different age groups shall achieve better performance on more readable texts, it is interesting to look at the differences in performance between CLM and RLM perplexity measures trained on Wiki-normal, Wiki-simple and Wiki-balanced corpora. Results on the WeeBit corpus are hard to interpret, since all perplexity measures show a weak negative correlation with the readability. On the OneStopEnglish corpus, both Wiki-simple perplexity measures perform the best, while on the Newsela corpus, RLM perplexity-balanced outperforms RLM perplexity-simple by 0.042 and CLM perplexity-simple outperforms CLM perplexity-balanced by 0.055. Both Wiki-normal perplexity measures are outperformed by a large margin by Wiki-simple and Wiki-balanced perplexity measures on the OneStopEnglish and the Newsela corpora. Similar observations can be made in regards to RSRS, which also leverages language model statistics. On all corpora Wiki-simple RSRS measures outperform Wiki-balanced RSRS measures and Wiki-balanced RSRS consistently outperforms Wiki-normal RSRS measures.

These results are not entirely compatible with our initial expectations that Wiki-balanced measures would be the most correlated with readability in most cases. On the other hand, the differences in performance between Wiki-balanced and Wiki-simple measures are not large and the positive correlation between readability and perplexity measures on the Newsela and OneStopEnglish corpora are quite strong which supports the hypothesis that more complex language structures and vocabularies of less readable texts would result in higher perplexity on these texts. According to our results, this phenomenon might not be very strong and only works if the training set is balanced in terms of readability classes for different ages. On the other hand, if the training set contains more texts for adults than for children, as in the case of language models trained just on the Wiki-normal corpus (and also BERT), this phenomenon disappears or even gets reverted, since language models trained on more complex language structures learn how to handle these difficulties.

The low performance of perplexity measures suggests that discourse cohesion and background knowledge leveraged by language models are not good indicators of readability and should therefore not be used in the readability formulas in the direct form. However, the results of CLM RSRS and RLM RSRS suggest that language models contain quite useful information when combined with other shallow lexical sophistication indicators. For English, the RLM RSRS-simple and the CLM RSRS-simple rank first with the average rank of 4.0. The CLM RSRS-balanced and RLM RSRS-balanced are the second best with the average rank of 5.0, together with the baseline average sentence length measure. CLM RSRS and RLM RSRS on Slovenian corpus also perform well with CLM RSRS-balanced being ranked first and RLM RSRS-balanced being the third. On the other hand, BERT RSRS is not well correlated with readability, with an average rank of 15.3 on the English corpora and the rank of 12.0 on the Slovenian corpus. This is not surprising, since all BERT perplexities are negatively correlated with the readability classes.

When it comes to cross-language transferability of readability measures (see column Diff. in Table 4), the most consistent ranking by performance is achieved by the RLM RSRS-balanced with no difference in ranking on English and Slovenian corpora. CLM RSRS-balanced, the best ranked measure on Slovenian corpus, also performs quite consistently with the difference in ranks of 2.0. Among the traditional measures, GFI presents the best balance in performance and consistency, ranking sixth on English and fourth on Slovenian. On the other hand, SMOG, which ranked very well on Slovenian (rank 2.0), ranked eleventh on English, which is the largest difference in ranking among all measures. The opposite can be said about the simplest readability measure, the average sentence length, which performed well on English (rank 3.0) and badly on Slovenian (rank 7.0).

To sum up, compared to perplexity scores and traditional readability measures, the proposed RSRS scores outperformed other scores on 2 out of 4 gold-standard datasets (see Table 3), achieved the best ranks, and showed the most stable cross-language performance (see Table 4).

5. Supervised neural approach

As mentioned in Section 2.5, recent trends in text classification show the domination of deep learning approaches which internally employ automatic feature construction. Surprisingly, even the most recent approaches to readability classification rely on hand crafted features and standard machine learning classifiers (Vajjala and Lucic 2018; Xia, Kochmar, and Briscoe 2016). In this Section, we describe how different types of neural classifiers can predict text readability and evaluate their performance.

The Section is divided into Section 5.1, where we describe the methodology, and Section 5.2, where we present the experimental scenario and the results of the conducted experiments.

5.1 Methodology

There exist several successful architectures of neural networks. We tested three distinct neural network approaches to text classification described in Section 2.5:

- Bidirectional Long short-term memory network (BiLSTM). We use the RNN approach proposed by Conneau et al. (2017) for classification. The bidirectional LSTM layer is a concatenation of forward and backward LSTM layers that read documents in two opposite directions. The max and mean pooling are applied to the LSTM output feature matrix in order to get the maximum and average values of the matrix. The resulting vectors are concatenated and fed to a linear layer responsible for producing final predictions.
- Hierarchichal attention networks (HAN). We use the identical architecture in this classifier as the one described in Yang et al. (2016) that takes hierarchical structure of text into an account through the two level attention mechanism (Bahdanau, Cho, and Bengio 2014; Xu et al. 2015) applied to word and sentence representations encoded by bidirectional LSTMs.
- Transfer learning. We use a pretrained BERT transformer architecture with 12 layers of size 768 and 12 self-attention heads. A linear classification head was added on top of the pretrained language model and the whole classification model was fine-tuned on every data set for 3 epochs. For English data sets, a pretrained uncased language model trained on BooksCorpus (800M words) (Zhu et al. 2015) and English Wikipedia (2,500M words) was used, while for the Slovenian school books corpus, a multi-lingual uncased language model trained on Wikipedia dumps of

hundred languages with the biggest Wikipedias was used (Devlin et al. 2018)[8].

We randomly split the Newsela and Slovenian school books corpora into a train (80% of the corpus), validation (10% of the corpus) and test (10% of the corpus) sets. Due to the small number of documents in OneStopEnglish and WeeBit corpora (see description in Section 3), we used five-fold cross validation on these corpora to get more reliable results. For every fold, the corpora were split into the train (80% of the corpus), validation (10% of the corpus) and test (10% of the corpus) sets.

BiLSTM and HAN classifiers were trained on the train set and tested on the validation set after every epoch (for a maximum of 100 epochs), and the best performing model on the validation set was selected as the final model and produced predictions on the test sets. The validation sets were also used in a grid search to find the best hyperparameters of the models. For BiLSTM, all combinations of the following hyperparameter values were tested before choosing the best combination, which is written in bold in the list below:

- Batch size: **8**, 16, 32
- Learning rates: 0.00005, **0.0001**, 0.0002, 0.0004, 0.0008
- Word embedding size: 100, **200**, 400
- LSTM layer size: **128**, 256
- Number of LSTM layers: 1, **2**, 3, 4
- Dropout after every LSTM layer: 0.2, **0.3**, 0.4

For HAN, all combinations of the following hyperparameter values were tested (the best combination is written in bold in the list below):

- Batch size: 8, **16**, 32
- Learning rates: 0.00005, **0.0001**, 0.0002, 0.0004, 0.0008
- Word embedding size: 100, **200**, 400
- Sentence embedding size: **100**, 200, 400

We used the same configuration for all the corpora and performed no corpus specific tweaking of classifier parameters. We measured the performance of all the classifiers in terms of accuracy (in order to compare their performance to the performance of the classifiers from the related work), weighted average precision, weighted average recall, and weighted average F-score. We calculate the weighted average precision, weighted average recall, and weighted average F-score by first calculating the precision (p_i) and recall (r_i) for each class i according to the following formulae:

$$p_i = \frac{TP_i}{TP_i + FP_i}$$

[8] Models are available at https://github.com/google-research/bert

$$r_i = \frac{TP_i}{TP_i + FN_i}$$

TP_i are true positive predictions (documents correctly classified into class i), FP_i are false positive predictions (documents incorrectly classified into class i), and FN_i are false negative predictions (documents incorrectly classified into other classes instead of class i). The weighted average precision (P_w) and weighted average recall (R_w) are defined with the following equations:

$$P_w = \frac{\sum_{i=1}^{n}(p_i * |c_i|)}{\sum_{i=1}^{n}|c_i|}$$

$$R_w = \frac{\sum_{i=1}^{n}(r_i * |c_i|)}{\sum_{i=1}^{n}|c_i|}$$

Given a corpus with readability classes $c_i, i \in [1, n]$, the precision p_i for class i is weighted with the number of documents belonging to that readability class ($|c_i|$). The same weighting scheme is used in a calculation of the weighted recall, where the recall r_i for the class i is weighted with the number of documents belonging to that readability class ($|c_i|$). The weighted average F-score is calculated as a weighted harmonic mean between P_w and R_w according to the following formula:

$$F1_w = 2 * \frac{P_w * R_w}{P_w + R_w}$$

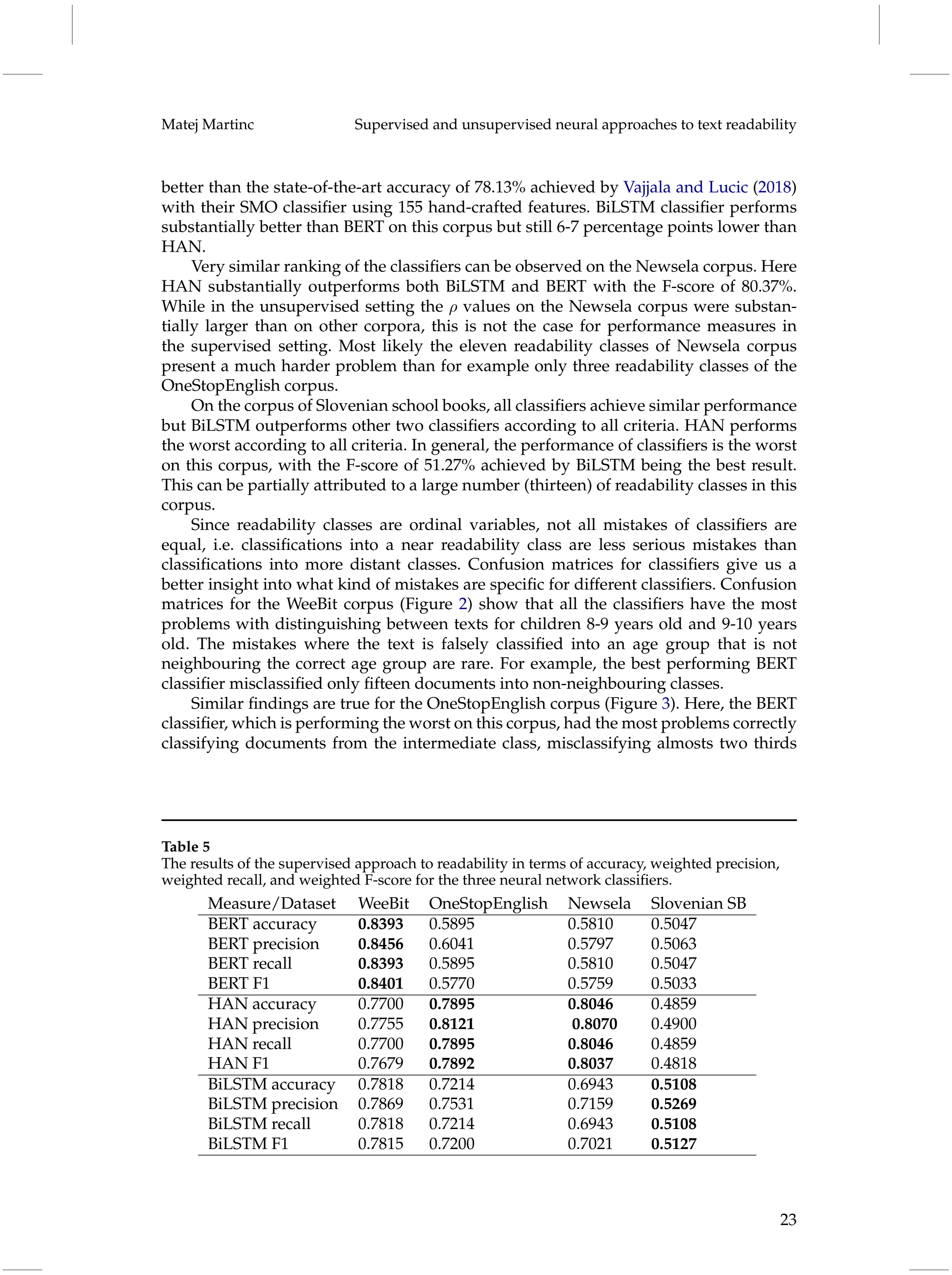
5.2 Experimental results

The results of supervised readability assessment using different architectures of deep neural networks are presented in Table 5.

On the WeeBit corpus, by far the best performance according to all measures was achieved by BERT. In terms of accuracy, BERT outperforms the second best BiLSTM by almost 6 percentage points, achieving the accuracy of 83.93%. HAN performs the worst on the WeeBit corpus according to all measures. BERT also outperforms the best reported accuracy from the literature reported by Xia, Kochmar, and Briscoe (2016) using the five-fold cross validation setting. By achieving 80.3%, it is better by about 4.5% percentage points.

On the other hand, BERT performs poorly on the OneStopEnglish and Newsela corpora. On both corpora, it is outperformed by the best performing classifier (HAN) by about 20 percentage points according to all criteria. We suspect that the main reason for the bad performance of BERT on these two corpora is the semantic similarity between classes. In these two corpora, the simplified versions of the original texts contain the same message as the original texts, but written in a more simplistic way. The results of our experiments suggest that because BERT is pretrained as a language model, it tends to rely more on semantic than structural differences during the classification phase and therefore performs better on problems with distinct semantic differences between readability classes. This is the case with the WeeBit and Slovenian school books corpora but not with the OneStopEnglish and Newsela corpora.

The best performance on the OneStopEnglish corpus is achieved by the HAN classifier with the accuracy of 78.95% in the five-fold cross validation setting. This is slightly



better than the state-of-the-art accuracy of 78.13% achieved by Vajjala and Lucic (2018) with their SMO classifier using 155 hand-crafted features. BiLSTM classifier performs substantially better than BERT on this corpus but still 6-7 percentage points lower than HAN.

Very similar ranking of the classifiers can be observed on the Newsela corpus. Here HAN substantially outperforms both BiLSTM and BERT with the F-score of 80.37%. While in the unsupervised setting the ρ values on the Newsela corpus were substantially larger than on other corpora, this is not the case for performance measures in the supervised setting. Most likely the eleven readability classes of Newsela corpus present a much harder problem than for example only three readability classes of the OneStopEnglish corpus.

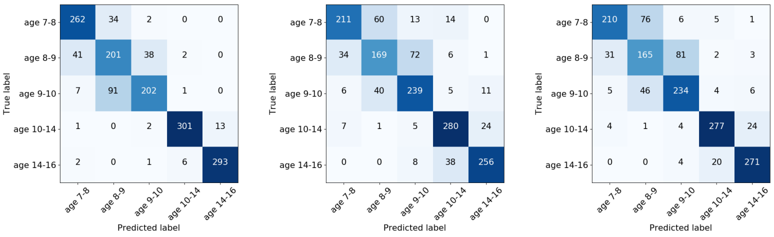
On the corpus of Slovenian school books, all classifiers achieve similar performance but BiLSTM outperforms other two classifiers according to all criteria. HAN performs the worst according to all criteria. In general, the performance of classifiers is the worst on this corpus, with the F-score of 51.27% achieved by BiLSTM being the best result. This can be partially attributed to a large number (thirteen) of readability classes in this corpus.

Since readability classes are ordinal variables, not all mistakes of classifiers are equal, i.e. classifications into a near readability class are less serious mistakes than classifications into more distant classes. Confusion matrices for classifiers give us a better insight into what kind of mistakes are specific for different classifiers. Confusion matrices for the WeeBit corpus (Figure 2) show that all the classifiers have the most problems with distinguishing between texts for children 8-9 years old and 9-10 years old. The mistakes where the text is falsely classified into an age group that is not neighbouring the correct age group are rare. For example, the best performing BERT classifier misclassified only fifteen documents into non-neighbouring classes.

Similar findings are true for the OneStopEnglish corpus (Figure 3). Here, the BERT classifier, which is performing the worst on this corpus, had the most problems correctly classifying documents from the intermediate class, misclassifying almosts two thirds

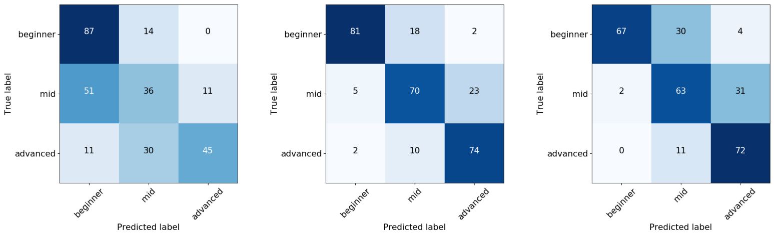
Table 5
The results of the supervised approach to readability in terms of accuracy, weighted precision, weighted recall, and weighted F-score for the three neural network classifiers.

Measure/Dataset	WeeBit	OneStopEnglish	Newsela	Slovenian SB
BERT accuracy	**0.8393**	0.5895	0.5810	0.5047
BERT precision	**0.8456**	0.6041	0.5797	0.5063
BERT recall	**0.8393**	0.5895	0.5810	0.5047
BERT F1	**0.8401**	0.5770	0.5759	0.5033
HAN accuracy	0.7700	**0.7895**	**0.8046**	0.4859
HAN precision	0.7755	**0.8121**	**0.8070**	0.4900
HAN recall	0.7700	**0.7895**	**0.8046**	0.4859
HAN F1	0.7679	**0.7892**	**0.8037**	0.4818
BiLSTM accuracy	0.7818	0.7214	0.6943	**0.5108**
BiLSTM precision	0.7869	0.7531	0.7159	**0.5269**
BiLSTM recall	0.7818	0.7214	0.6943	**0.5108**
BiLSTM F1	0.7815	0.7200	0.7021	**0.5127**



a) BERT b) HAN c) BiLSTM

Figure 2
Confusion matrices for BERT, HAN, and BiLSTM on the WeeBit corpus.



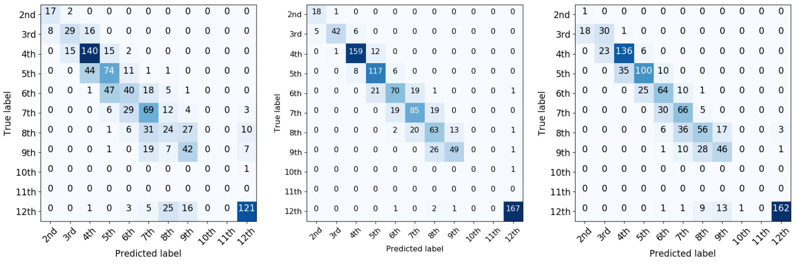
a) BERT b) HAN c) BiLSTM

Figure 3
Confusion matrices for BERT, HAN, and BiLSTM on the OneStopEnglish corpus.

of the documents. HAN and BiLSTM classifiers performed better, both misclassifying about one third of the documents from the intermediate class. Both classifiers had the least problems with documents from the advanced class, misclassifying approximately 15% of these documents.

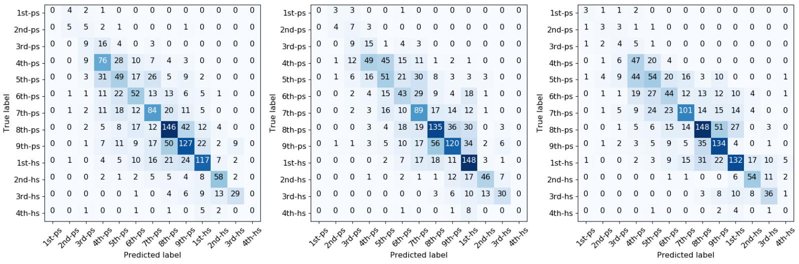
Confusion matrices for the Newsela corpus (Figure 4) follow a similar pattern, even though the number of classes is much larger and classes are unbalanced. Unsurprisingly, no classifier predicted any documents to be in two minority classes (10th and 11th grade) with very little training examples. The confusion matrix of the BERT classifier also clearly shows that this classifier has problems on this dataset, since the false predictions are more dispersed across classes than in the case of HAN and BiLSTM which classified a large majority of misclassified instances into neighbouring classes. The most visible error made by BERT is misclassifying 50 documents from the 12th grade into non-neighbouring classes. On the other hand, the best performing HAN classifier misclassified only four examples from the 12th grade and altogether misclassified only eleven examples into non-neighbouring classes.

Confusion matrices for the Slovenian school books corpus (Figure 5) are similar, which is unsurprising, provided that all classifiers achieved similar performance on this dataset. The biggest spread of misclassified documents is visible for the classes in



a) BERT b) HAN c) BiLSTM

Figure 4
Confusion matrices for BERT, HAN, and BiLSTM on the Newsela corpus.



a) BERT b) HAN c) BiLSTM

Figure 5
Confusion matrices for BERT, HAN, and BiLSTM on the Slovenian school books corpus.

the middle of readability range (from the 4th grade primary school to the 1st grade high school). Even though F-score results are relatively low on this dataset for all classifiers (the best F-score of 51.27% was achieved by BERT), all confusion matrices clearly show that a majority of misclassified examples were put into classes close to the correct one, suggesting that classification approaches to readability prediction can also be reliably used for Slovenian.

Overall, the classification results suggest that neural networks are a viable option for the supervised readability prediction,. Our approach managed to outperform all standard machine learning classifiers, leveraging extensive feature engineering (Xia, Kochmar, and Briscoe 2016; Vajjala and Lucic 2018), on both corpora, where comparisons are available.

6. Conclusion

We presented a set of novel approaches for determining readability of documents using deep neural networks. This is, to the best of our knowledge, the first attempt to leverage neural language models and neural network classifiers for readability prediction. The approaches are tested on a number of manually labeled English and Slovenian corpora.

We improve the performance over current state-of-the-art approaches to readability prediction in both unsupervised and supervised settings.

The results suggest that unsupervised approaches to readability prediction that only take background knowledge and discourse cohesion into account cannot compete with the approaches based on shallow lexical sophistication indicators (e.g., sentence length, word length, etc.). However, combining the components of several readability indicators into the new RSRS (ranked sentence readability score) measure does improve the correlation with true readability scores. Additionally, the RSRS measure is adaptable, robust, and transferable across languages.

The functioning of the proposed RSRS measure can be customized and influenced by the choice of the training set. This is a desired property, since it enables personalization and localization of the readability measure according to the educational needs, language, and topic. The usability of this feature might be limited for under-resourced languages, since sufficient amount of documents needed to train a language model that can be used for the task of readability prediction in a specific customized setting might not be available. On the other hand, our experiments on the Slovenian language show, that a relatively small 2.4 million word training corpora for language models is sufficient to outperform traditional readability measures.

The results of the unsupervised approach to readability prediction on the corpus of Slovenian school books are not entirely consistent with the results reported by the previous Slovenian readability study (Škvorc et al. 2018), where the authors reported that simple indicators of readability, such as average sentence length, performed quite well. Our results show that the average sentence length performs very competitively on English but ranks badly on Slovenian. This inconsistency in results might be explained with the difference in corpora used for the evaluation of our approaches. While Škvorc et al. (2018) conducted experiments on a corpus of magazines for different age groups (which we used for language model training), our experiments were conducted on a corpus of school books, which contains school books for sixteen distinct school subjects with very different topics ranging from literature, music and history to math, biology and chemistry. This might hint that the variance in genres and covered topics has an important effect on the ranking and performance of different readability measures. Further experiments on other Slovenian datasets, which we plan to conduct in the future, are required to confirm this hypothesis.

In the supervised approach to determining readability, we show that neural classifiers outperform state-of-the-art standard approaches on both corpora (WeeBit and OneStopEnglish) where comparison is available. However, the performance of different classifiers varies across different corpora, which is especially true for the BERT classifier. We hypothesize that this is due to its language model pretraining with focus on language understanding tasks, which makes the classifier sensitive to semantic information and therefore not appropriate for distinguishing between documents from different readability classes with similar meaning. More consistent behaviour is achieved by the HAN classifier that manages to outperform state-of-the-art approach proposed by Vajjala and Lucic (2018) on the OneStopEnglish corpus. Experiments also show that the attention based HAN classifier might be more appropriate for readability classification than the BiLSTM classifier, most likely due to more comprehensive context information. Even though BiLSTM slightly outperforms HAN on two out of four corpora, it is surpassed by a large margin on the other two corpora by HAN. These two corpora are OneStopEnglish and Newsela, where documents from different readability classes are semantically similar, which suggests that HAN classifier might be better capable of leveraging syntactic and structural information and relies less on semantic differences.

The differences in performance between classifiers on different corpora suggest that tested classifiers take different types of information into account. Provided this hypothesis is correct, some gains in performance might be achieved if these classifiers are combined. We plan to test a neural ensemble approach for the task of predicting readability in the future.

A more detailed look into confusion matrices of all classifiers on all corpora shows that the most common mistake all classifiers make is to misclassify a document into a neighbouring class. This makes our classification approaches to readability relatively informative and reliable even on the corpus of Slovenian school books, where the best F-score is relatively low compared to the very high results on the English corpora. The ordinal nature of readability classes will be further explored and exploited in the future work, when supervised (ordinal) regression approaches for determining readability will be tested.

We also plan to test the cross-genre and cross-language transferability of the proposed supervised and unsupervised approaches. This requires new readability datasets for different languages and genres which are currently rare or not publicly available. This might open opportunity to further improve the proposed unsupervised readability score.

Acknowledgments

The research was financially supported by the European social fund and Republic of Slovenia, Ministry of Education, Science and Sport through project Quality of Slovene textbooks (KaUČ). The work was supported by the Slovenian Research Agency (ARRS) core research programmes P6-0411 and P2-0103 and the project Terminology and knowledge frames across languages (J6-9372). This work has also received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825153 (EMBEDDIA). The results of this publication reflect only the authors' views and the EC is not responsible for any use that may be made of the information it contains.

References

Anderson, Jonathan. 1981. Analysing the readability of english and non-english texts in the classroom with lix. In *Seventh Australian Reading Association Conference*, pages 1–12, ERIC.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bai, Shaojie, J Zico Kolter, and Vladlen Koltun. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

Bengio, Yoshua, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166.

Bird, Steven and Edward Loper. 2004. Nltk: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 31, Association for Computational Linguistics.

Chen, Stanley F and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394.

Collins-Thompson, Kevyn and Jamie Callan. 2005. Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, 56(13):1448–1462.

Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Conneau, Alexis, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Conneau, Alexis, Holger Schwenk, Loïc Barrault, and Yann Lecun. 2016. Very deep convolutional networks for text classification. *arXiv preprint arXiv:1606.01781*.

Crossley, Scott A, Stephen Skalicky, Mihai Dascalu, Danielle S McNamara, and Kristopher Kyle. 2017. Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes*, 54(5-6):340–359.

Dale, Edgar and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Davison, Alice and Robert N Kantor. 1982. On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading research quarterly*, pages 187–209.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Council of Europe, Council for Cultural Co-operation. Education Committee. Modern Languages Division. 2001. *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge University Press.

Goldberg, Yoav and Jon Orwant. 2013. A dataset of syntactic-ngrams over time from a very large corpus of english books. In *Second Joint Conference on Lexical and Computational Semantics*, pages 241–247.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.

Gunning, Robert. 1952. *The technique of clear writing*. McGraw-Hill, New York.

Howard, Jeremy and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Kandel, Lilian and Abraham Moles. 1958. Application de l'indice de flesch à la langue française. *Cahiers Etudes de Radio-Télévision*, 19(1958):253–274.

Kim, Yoon, Yacine Jernite, David Sontag, and Alexander M Rush. 2016. Character-aware neural language models. In *AAAI*, pages 2741–2749.

Kincaid, J Peter, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Institute for Simulation and Training, University of Central Florida.

Kusner, Matt, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966.

Logar, Nataša, Miha Grčar, Marko Brakus, Tomaž Erjavec, Špela Arhar Holdt, Simon Krek, and Iztok Kosem. 2012. *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Trojina, zavod za uporabno slovenistiko.

Mc Laughlin, G Harry. 1969. Smog grading - a new readability formula. *Journal of reading*, 12(8):639–646.

Mikolov, Tomáš, Anoop Deoras, Stefan Kombrink, Lukáš Burget, and Jan Černockỳ. 2011. Empirical evaluation and combination of advanced language modeling techniques. In *Twelfth Annual Conference of the International Speech Communication Association*.

Petersen, Sarah E and Mari Ostendorf. 2009. A machine learning approach to reading level assessment. *Computer speech & language*, 23(1):89–106.

Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Schwarm, Sarah E and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 523–530, Association for Computational Linguistics.

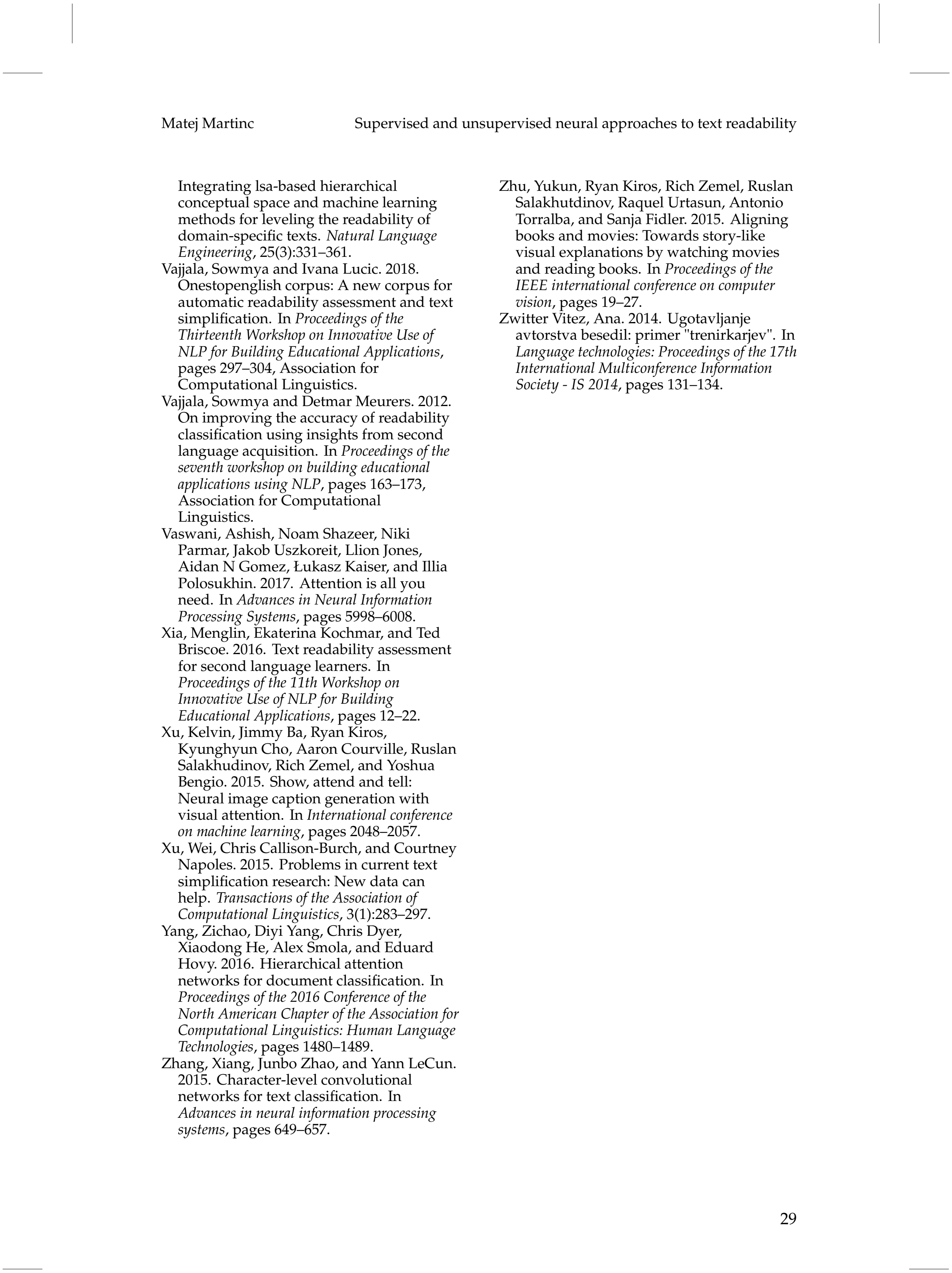
Si, Luo and Jamie Callan. 2001. A statistical model for scientific readability. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 574–576, ACM.

Škvorc, Tadej, Simon Krek, Senja Pollak, Špela Arhar Holdt, and Marko Robnik-Šikonja. 2018. Evaluation of statistical readability measures on slovene texts. In *Conference on Language Technologies and Digital Humanities*, pages 240–247, Ljubljana University Press, Faculty of arts.

Smith, Edgar A and R.J. Senter. 1967. Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories (US)*, pages 1–14.

Tang, Duyu, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1422–1432.

Tseng, Hou-Chiang, Berlin Chen, Tao-Hsing Chang, and Yao-Ting Sung. 2019.



Integrating lsa-based hierarchical conceptual space and machine learning methods for leveling the readability of domain-specific texts. *Natural Language Engineering*, 25(3):331–361.

Vajjala, Sowmya and Ivana Lucic. 2018. Onestopenglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, Association for Computational Linguistics.

Vajjala, Sowmya and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the seventh workshop on building educational applications using NLP*, pages 163–173, Association for Computational Linguistics.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Xia, Menglin, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22.

Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.

Xu, Wei, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association of Computational Linguistics*, 3(1):283–297.

Yang, Zichao, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.

Zhang, Xiang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.

Zhu, Yukun, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

Zwitter Vitez, Ana. 2014. Ugotavljanje avtorstva besedil: primer "trenirkarjev". In *Language technologies: Proceedings of the 17th International Multiconference Information Society - IS 2014*, pages 131–134.

Appendix C: Predicting Slovene Text Complexity Using Readability Measures

Predicting Slovene Text Complexity Using Readability Measures

Tadej Škvorc[*]
Simon Krek[**]
Senja Pollak[***]
Špela Arhar Holdt[****]
Marko Robnik-Šikonja[*****]

Cobiss type: 1.01
UDC: 003.295:821.163.6

IZVLEČEK

NAPOVEDOVANJE KOMPLEKSNOSTI SLOVENSKIH BESEDIL Z UPORABO MER BERLJIVOSTI

[1] *Večina obstoječih formul za merjenje berljivosti je zasnovana za besedila v angleškem jeziku, na katerih je tudi ocenjena njihova kakovost. V našem članku predstavimo prilagoditev izbranih mer za slovenščino. Uspešnost desetih znanih formul ter osmih dodatnih kriterijev berljivosti ocenimo na petih skupinah besedil: otroških revijah, splošnih revijah, časopisih, tehničnih revijah in zapisnikih sej državnega zbora. Te skupine besedil imajo različne ciljne publike, zaradi česar predpostavimo, da uporabljajo različne stile pisanja, ki bi jih formule in kriteriji berljivosti morali zaznati. V analizi pokažemo, katere formule in kriteriji berljivosti delujejo dobro in s katerimi razlik med skupinami nismo mogli zaznati.*

[2] *Ključne besede: berljivost, obdelava naravnega jezika, analiza besedil*

ABSTRACT

[1] *The majority of existing readability measures are designed for English texts. We aim to adapt and test the readability measures on Slovene. We test ten well-known readability formulas and eight additional readability criteria on five types of texts: children's magazines, general magazines, daily newspapers, technical magazines, and transcriptions of national assembly sessions. As these groups of texts target different audiences, we assume that the differences in writing styles should be reflected in their readability scores. Our analysis shows which readability measures perform well on this task and which fail to distinguish between the groups.*

[2] *Keywords: readability, natural language processing, text analysis*

1. Introduction

[1] In English, the problem of determining text readability (i.e. how easy a text is to understand) has long been a topic of research, with its origins in the 19th century (Sherman 1893). Since then, many different methods and readability measures have been developed, often with the goal of determining whether a text is too difficult for its target age group. Even though the question of readability is complex from a linguistic standpoint, a large majority of existing measures are based on simple heuristics. There has been little research on readability of languages other than English, therefore we aim to apply these measures to

Slovene and evaluate how well they perform.

[2] There are several factors that might cause these measures to perform poorly on non-English languages, such as:

- Many measures are fine-tuned to correspond to the grade levels of the United States education system. It is likely a different fine-tuning would be needed for other languages, as a.) their education system is different from the US system, and b.) the differences in readability between grade levels are likely to be different between languages, meaning that each language would require specifically tuned parameters.
- Some measures utilize a list of common English words and their results depend on the definition of this list. For Slovene, there currently does not exist a publicly available list of common words, so it is not known how such measures would perform.
- The existing readability measures do not use the morphological information to determine difficult words but rely on syllable and character counts, or a list of difficult words. As Slovene is morphologically much more complex than English, words with complex morphology are harder to understand than those with simple morphology, even if they have the same number of characters or syllables.

[3] We analyze the commonly used readability measures (as well as some novel measures) on Slovene texts and propose a word list needed to implement the word-list-based measures. We calculate statistical distributions of scores for each readability measure across subcorpora and assess the ability of measures to distinguish between different subcorpora using a variety of statistical tests. We show that machine learning classification models, using a combination of readability measures, can predict the subcorpus a given text belongs to.

[4] The paper extends the short version of the paper presented in Škvorc et al. (2018) and is structured as follows. We first present the related work on readability measures and describe the readability measures used in our analysis. The methodology of the analysis is presented next, followed by the results split into three sections. The last section concludes the paper and presents ideas for further work.

2. Related Work

[1] For English, there exists a variety of works focused on determining readability by using readability formulas. Those formulas rely on different features of the text such as the average sentence length, percentage of difficult words, and the average number of characters per word. Examples of such measures include the Coleman-Liau index (Coleman and Liau 1975), LIX (Björnsson 1968), and the automated readability index (ARI) (Senter and Smith 1967). Some formulas, like the Flesch-Kincaid grade level (Kincaid et al. 1975) and SMOG (Mc Laughlin 1969) use the number of syllables per word to determine if a word is difficult. Additionally, some measures (e.g., the Spache readability formula (Spache 1953) and Dale-Chall readability formula (Dale and Chall 1948) rely on a pre-constructed list of difficult words.

[2] Aside from the readability formulas, there exists a variety of other approaches that can be used to determine readability (Bailin and Grafstein 2016). For example, various machine-learning approaches can be used to obtain better results than readability formulas, such as the approach presented in Francois and Miltsakaki (2012), which outperforms readability formulas on French text.

[3] There is little work attempting to apply these measures to Slovene texts. Most work dealing with the readability of Slovene text is focused on manual methods. For example, Justin (2009) analyzes Slovene textbooks from a variety of angles, including readability. On the other hand, works that focus on automatic readability measures are rare. Zwitter Vitez (2014) uses a variety of readability measures for author recognition in Slovene text, but we found no works that used them to determine readability.

[4] In addition to Slovene, some related works evaluate readability measures on other languages. Debowski et al. (2015) evaluate readability formulas on Polish text and show that they obtain better results by using a more complex, machine-learning-based approach.

3. Readability Measures

[1] In our analysis, we used two groups of readability measures:

- **Existing readability formulas for English:** we focused mainly on popular methods that have been shown to achieve good results on English texts. These measures mostly rely on easy-to-obtain features such as a number of difficult words, sentence length, and word length.
- **Natural-language-processing-based readability criteria:** we used additional criteria that are not present in the existing readability formulas but can be obtained from tools for automatic language processing, such as the percentage of verbs, number of unique words, and morphological difficulty of words. In the existing English formulas, such criteria are not used but they might contain useful information for determining the readability of Slovene texts.

[2] In the following two subsections we present the established readability measures for grading English text and our proposed additional criteria.

4. Existing Readability Formulas

[1] There exists a variety of ways to measure the readability of texts written in English. For our analysis, we used 10 readability formulas given below. The entities used in the expressions correspond to the number of occurrences of a given entity, e.g., word corresponds to the number of words in a measured text.

- **Gunning fog index** (Gunning 1952) is calculated as:

$$\mathrm{GFI} = 0.4\frac{\text{words}}{\text{sentences}} + 100\frac{\text{complex words}}{\text{words}},$$

  where a word is considered complex if it contains three or more syllables. As there exists no established automatic method for counting syllables of Slovene words, we used a rule-based approach designed for English. The resulting score is calibrated to the grade level of the USA education system.

- **Flesch reading ease** (Kincaid et al. 1975) is calculated as:

$$\mathrm{FRE} = 206.835 - 1.015\frac{\text{words}}{\text{sentences}} - 84.6\frac{\text{syllables}}{\text{words}}.$$

  The score does not correspond to grade levels. Instead, the higher the value, the easier the text is considered to be. A text with a score of 100 should be easily understood by 11-year-old students, while a text with a score of 0 should be intended for university graduates.

- **Flesch–Kincaid grade level** (Kincaid et al. 1975) is similar to Flesch reading ease, but does correspond to grade levels. It is calculated as:

$$\mathrm{FKGL} = 0.39\frac{\text{words}}{\text{sentences}} + 11.8\frac{\text{syllables}}{\text{words}} - 15.59.$$

- **Dale–Chall readability formula** (Dale and Chall 1948) is calculated as:

$$\mathrm{DCRF} = 0.1579\frac{\text{difficult words}}{\text{words}} + 0.0496\frac{\text{words}}{\text{sentences}}.$$

  [1] The formula requires a predefined list of common (easy) words and the words which are not on the list are considered as difficult. The novelty of the Dale-Chall Formula was that it did not use word-length counts but a count of “hard” words which do not appear on a specially designed list of common words. This list was defined as the words familiar to most of the 4th-grade students: when 80 percent of the fourth-graders indicated that they knew a word, the word was added to the list.

  [2] Higher scores indicate that the text is harder, but the resulting score does not correspond to grade levels, nor is it appropriate for text aimed at children below 4th grade. In our analysis, we obtained the difficult words in two ways:

  1. By constructing a list of “easy” words and considering every word not on the list as difficult. The list of easy words is described later in the paper.
  2. By considering words with more than seven characters as difficult.

- **Spache readability formula** (Spache 1953) is calculated as:

$$\mathrm{SRF} = 0.141\frac{\text{words}}{\text{sentences}} + 8.6\frac{\text{unique difficult words}}{\text{unique words}} + 0.839.$$

  Difficult words are defined as words that do not appear in the list of commonly used words, which is the same as the one used in the Dale–Chall readability formula. This method was specifically designed for texts targeting children up to the fourth grade and was not designed to perform well on harder text. The obtained score corresponds to grade levels.

- **Automated readability index** (Senter and Smith 1967) is calculated as:

$$\text{ARI} = 4.71\frac{\text{characters}}{\text{words}} + 0.5\frac{\text{words}}{\text{sentences}} - 21.43.$$

The formula was designed so that it could be automatically captured in times when texts were written on typewriters and therefore it does not use information relating to syllables or difficult words. The obtained score corresponds to grade levels.

- **SMOG (Simple Measure of Gobbledygook)** (McLaughlin 1969) can be calculated as:

$$\text{SMOG} = 1.043\sqrt{\text{difficult words}\frac{30}{\text{sentences}}} + 3.1291,$$

where difficult words are defined as words with three or more syllables. The score corresponds to grade levels.

- **LIX** (Bjornsson 1968) is calculated as:

$$\text{LIX} = \frac{\text{words}}{\text{sentences}} + 100\frac{\text{long words}}{\text{words}},$$

where long words are defined as words consisting of more than six characters. LIX is the only measure we used that was not designed specifically for English but for a variety of languages. Because of this, it does not use syllables or a list of unique words. The score does not correspond to grade levels.

- **RIX** (Anderson 1983) is a simplification of LIX, and is calculated as:

$$\text{RIX} = \frac{\text{long words}}{\text{sentences}}.$$

- **Coleman-Liau index** (Coleman and Liau 1975) is calculated as:

$$\text{CLI} = 0.0588\text{L} - 0.296\text{S} - 15.8,$$

where L is the average number of letters per 100 words and S is the average number of sentences per 100 words. The obtained score corresponds to grade levels.

5. Language-Processing-Based Readability Criteria

[1] The readability formulas described in the previous section use a low number of common criteria, such as the number of syllables in words or the number of words in a sentence. In our analysis, we also analyzed Slovene texts using the following additional statistics:

- percentage of long words,
- percentage of difficult words,
- percentage of verbs,
- percentage of adjectives,
- percentage of unique words,

- average sentence length.

[2] Many of these (percentage of long words, difficult words, unique words, and average sentence length) are used as features in the readability measures described above. We evaluate them individually to determine how important each of them is for Slovene texts. The **percentage of verbs** is used because a higher number of verbs can indicate more complex sentences with multiple clauses. The **percentage of adjectives** was chosen because we assumed a higher percentage of adjectives could indicate longer, more descriptive sentences that are harder to understand.

[3] To take into account richer morphology of Slovene and a less fixed word order compared to English, we computed two additional criteria:

- **Context of difficult words**, which is the average number of difficult words that appear in a context (i.e. the three words before or after the word) of a difficult word. Difficult words are defined as words that do not appear on the list of common words. The intuition behind this metric is that a difficult word that appears in the context of easy words is easier to understand than if it is surrounded by other difficult words since its meaning can be more easily inferred from the context.
- **Average morphological difficulty**, where we use the Slovene morphological lexicon Sloleks (Arhar Holdt 2009) to assign a “morphological difficulty” score to each word. Sloleks is a lexicon of word forms and contains frequency information for morphological variants of over 100,000 lemmas (base forms of words as defined in a dictionary). We use the relative frequency of a word variant compared to other variants of the same lemma as the morphological difficulty score.

[4] In addition, we also calculated the number of words in each document, even if in our case, it cannot be interpreted as a criterion for determining readability since it is largely determined by the type of document. E.g., the documents belonging to the subcorpus of newspapers contain individual articles and are therefore short, while the subcorpus of computer magazines contains entire magazines which are considerably longer.

6. Analysis of Slovene Texts

[1] In this section, we describe the methodology used for our analysis. In the first subsection, we describe the data sets on which we conducted our analysis. In the second subsection, we describe how we constructed the list of easy words used in some of the readability measures.

6.1. Data Sets

[1] We created a set of subcorpora from the Gigafida reference corpus of written Slovene (Logar et al. 2012). Gigafida contains 39,427 Slovene texts released from 1990 to 2011, for a total of 1,187,002,502 words. We focused on texts published in magazines, newspapers, and books while ignoring texts collected from the internet. The texts in the Gigafida corpus are segmented into paragraphs and sentences, tokenized, and part-of-speech tagged using the Obeliks tagger (Grčar et al. 2012). We grouped the texts based on the intended audience, resulting in the following subcorpora:

- **Children's magazines** include magazines aimed at younger children (to be read independently or by their parents), namely Cicido and Ciciban.
- **Pop magazines** contain magazines aimed at the general public, namely Lisa, Gloss, and Stop.
- **Newspapers** contain general adult population newspapers, namely Delo and Dolenjski list.
- **Computer magazines** include magazines focusing on technical topics relating to computers, namely Monitor, Računalniške novice, PC & Mediji, and Moj Mikro.
- **National Assembly** includes transcriptions of sessions from the National Assembly of Slovenia.

[2] In Table 1 we show the number of documents in each subcorpus and the average number of words per document. The subcorpus of newspapers contains the largest number of documents, while the subcorpus of text sourced from the National Assembly of Slovenia contains the fewest.

Table 1: The number of documents and the average number of words per document for each subcorpus.

Subcorpus	**#docs**	**Avg. #words / doc**	**Total #words**
Children's magazines	125	5,488	686,000
Pop magazines	247	33,967	8,389,849
Newspapers	14,011	12,881	180,475,691
Computer magazines	163	110,875	18,072,625
National Assembly	35	58,841	2,059,435

[3] Our hypothesis is that the readability measures will be able to distinguish texts from different subcorpora. We assume that children's magazines will be easily distinguishable from other genres that are addressing an adult population. We also suppose that general magazines are less complex than specialized magazines. The National Assembly transcripts were included as they differ from other texts in two major ways: a.) they are transcripts of spoken language and b.) they relate to a highly technical subject matter. Because of this, we were interested in how readability measures would grade them. To test our hypothesis and to determine how well each readability measure works, we analyzed texts from each subcorpus to obtain a score distribution for each measure. The scores were calculated separately for each source text (e.g., one magazine article, a newspaper, or one assembly session).

6.2. List of Common Words

[1] For designing the list of common words, we took a corpus-based approach. Note that the methodology to create a list of common words from language corpora was already tested for other languages, (see e.g., Kilgarriff et al. 2014). We used four corpora to create a list of common words: Kres, Janes, Gos, and Šolar:

- **Šolar** (Kosem et al. 2011) contains 2,703 texts written by pupils in Slovenia from grades 6 to 13 (grade 6 to 9 in primary school, and grade 1 to 4 in secondary school). The texts include essays, summaries, and answers to examination questions.
- **Gos** (Verdonik et al. 2011) contains around 120 hours of recorded spoken Slovene (1,035,101 words), as well as transcriptions of the recordings. The recordings are collected from a variety of sources, including conversations, television, radio, and phone calls. Around 10% of the corpus consists of recorded lessons in primary and secondary schools.
- **Janes** (Fišer et al. 2014) contains Slovene texts from various internet sources, such as tweets, forum posts, blogs, comments, and Wikipedia talk pages.
- **Kres** (Logar Berginc and Šuster 2009) is a sub-corpus of Gigafida that is balanced with respect to the source (e.g. newspapers, magazines, or internet).

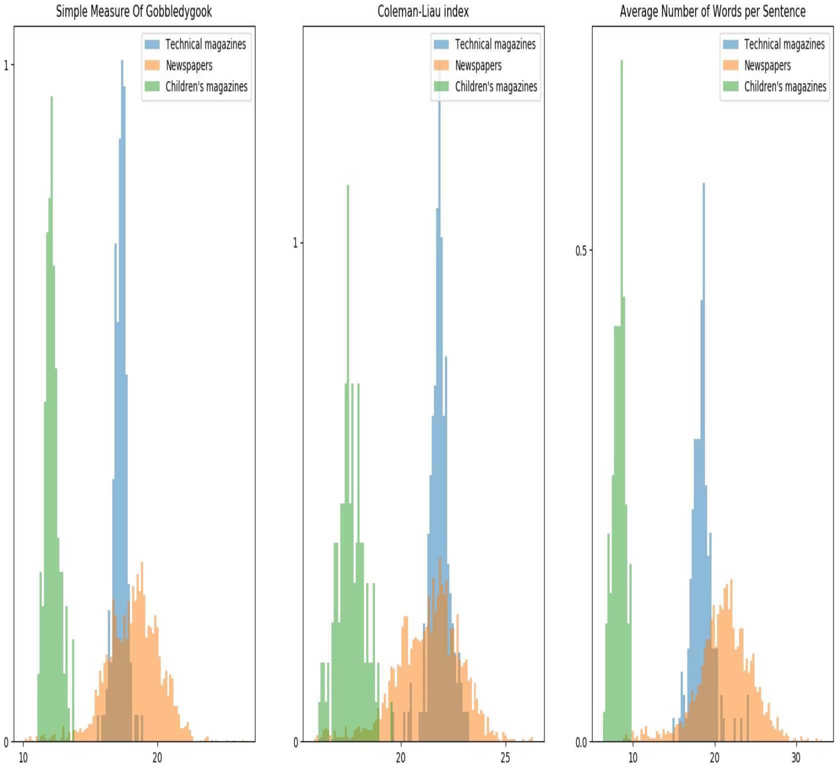
[2] We extracted the most common words and defined the common words as the ones that appear frequently in all four corpora (and are therefore not specific to a certain text type). We use four corpora to include texts that primarily reflect language production by different language users (Gos, Janes, Šolar), as well as texts that primarily reflect standard language (Kres). We aimed at covering younger school-going population (Šolar) and adults. For some corpora, we could have assigned words to different age levels (e.g. using pupils' grade levels in Šolar or using the age groups available in Gos metadata), but these corpora are very specific and the resulting word groups would mainly reflect the genre instead of age levels. Because of this, we opted for the approach of crossing the word lists to obtain a single list. The overlap of the most common words in four corpora eliminates frequent words which are typical for only one of the corpora (e.g. administrative language in Kres, spoken language markers in Gos, Twitter-specific usage in Janes, and literary references from essays in Šolar).

[3] From each corpus, we extracted the 10,000 most frequent word lemmas and part-of-speech tuples. In order to construct a list of common words representative of Slovene language, we selected the word lemmas that occurred in the most frequent word lists of all the four corpora. We obtained a list of 2,562 common words, which we used in readability measures.

7. Results

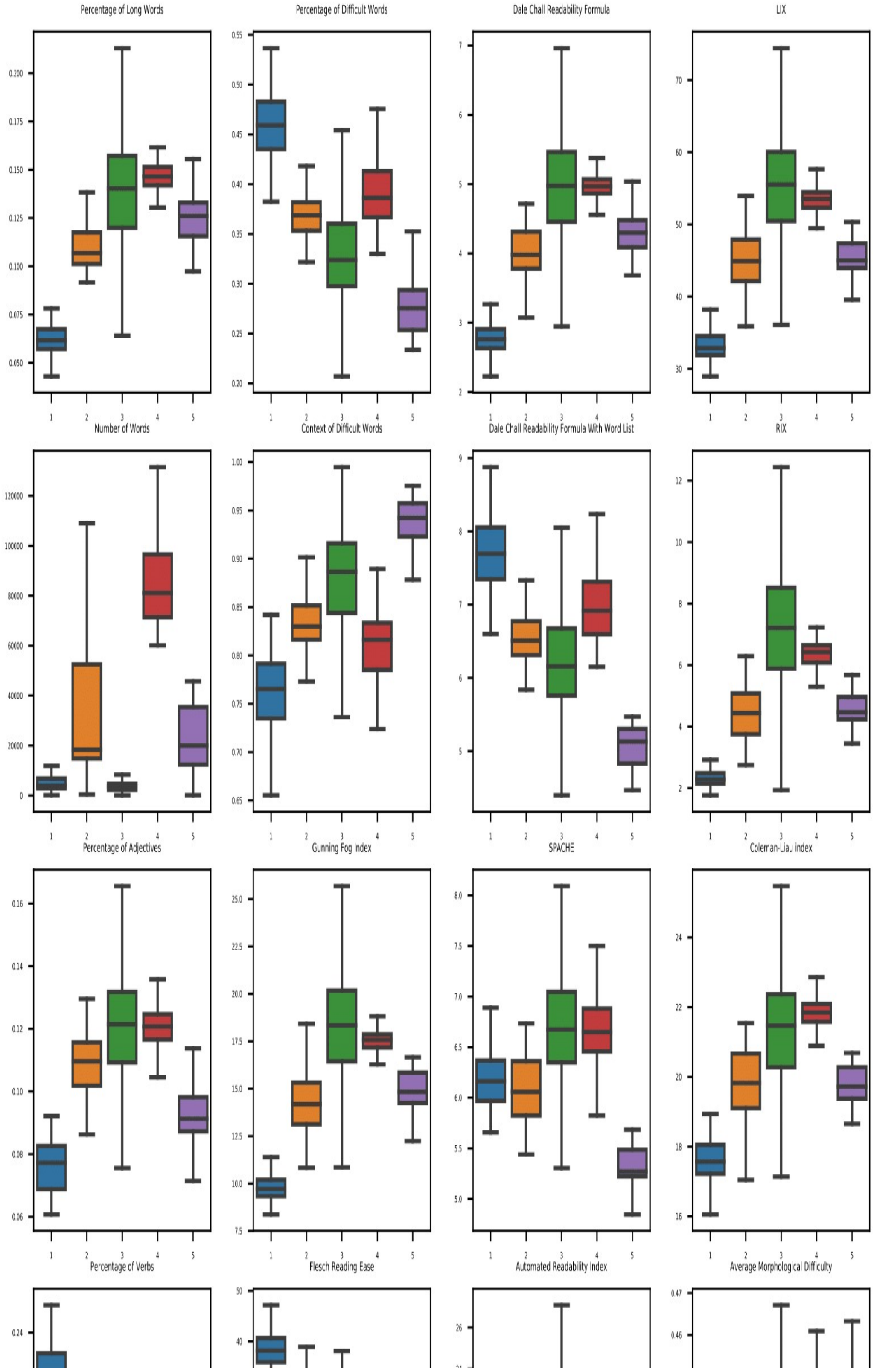
[1] For each text in each subcorpus, we calculated readability scores using all readability measures described in the previous section. In Figure 1 we present a few examples of obtained score distributions. We show distributions for three text subcorpora (children's magazines, newspapers, and technical magazines) and three readability scores (Goobledygook, Coleman-Liau, and the average number of words in a sentence).

Figure 1: The score distributions for three text subcorpora and three readability measures. The distributions show that technical magazines readability scores are the most consistent, while newspapers' scores are more diverse. Children's magazines' scores have a strong peak on the left-hand side (easier texts) that is well separated from the other sources.



[2] To show a compact overview of all included readability measures we calculated the median, first and third quartiles of the distribution for each score and each text subcorpus. The box-and-whiskers plots showing these results are visualized in Figure 2 which shows that most readability measures are able to distinguish between different subcorpora. Additionally, some of the readability measures confirm our original hypothesis, i.e. they are able to distinguish children's magazines from other genres that are addressing adult population, and evaluate general magazines as less complex than computer magazines.

Figure 2: The scores of each readability measure for each subcorpus of texts, represented with box plots. The subcorpora depicted from left to right are: 1.) Children's magazines, 2.) General magazines, 3.) Newspapers, 4.) Computer magazines, and 5.) National assembly transcriptions. The boxes show the first, second, and third quartile of the distributions while the whiskers extend for 1.5 IQR past the first and third quartile.


Percentage of Long Words
0.200
0.175
0.150
0.125
0.100
0.075
0.050
1
2
3
4
5
Percentage of Difficult Words
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
Dale Chall Readability Formula
7
6
5
4
3
2
LIX
70
60
50
40
30
Number of Words
120000
100000
80000
60000
40000
20000
0
Context of Difficult Words
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
Dale Chall Readability Formula With Word List
9
8
7
6
5
RIX
12
10
8
6
4
2
Percentage of Adjectives
0.16
0.14
0.12
0.10
0.08
0.06
Gunning Fog Index
25.0
22.5
20.0
17.5
15.0
12.5
10.0
7.5
SPACHE
8.0
7.5
7.0
6.5
6.0
5.5
5.0
Coleman-Liau index
24
22
20
18
16
Percentage of Verbs
0.24
Flesch Reading Ease
50
40
Automated Readability Index
26
Average Morphological Difficulty
0.47
0.46

[3] Figure 2 allows for an additional interpretation of readability measures. For example, children's magazines vs. general magazines vs. newspapers mean scores show increasing complexity in the following measures: Percentage of long words, Flesh Kincaid Grade Level, Gunning Fog Index, Dale-Chall Readability Formula (based on complexity defined by syllables), Context of Difficult Words, SMOG, LIX, RIX and Automated Readability Index. All these measures consider the length of words and/or sentences. The percentage of adjectives also seems to correlate with the complexity of these three text types, although to a lesser extent. The same holds for Flesh Reading Ease, since higher scores indicate lower complexity. For the majority of these measures, the distinction between newspapers and specialized computer magazines is either less evident or not evident at all, but they do indicate that computer magazines are less readable than general magazines.

[4] Scores using the list of common words do not lead to the same conclusions. Percentage of Difficult Words and Dale-Chall Readability Formula with word list do not reflect the complexity of genres, but to some extent, they do distinguish between general and specialized texts (i.e. newspapers and general magazines have lower scores than specialized computer magazines). One of the reasons for the relatively high scores for the complexity of children magazines might be in the large proportion of literary language, such as in poems for children with many words not in the list of common words. For example, “KRAH, KRAH, KRAH! MENE NIČ NI STRAH!” (Krah, krah, krah! I am not afraid!) has 7 words, out of which 4 are on the list of simple words, while the interjection KRAH is not on the simple words list. Therefore, the proportion of difficult words in this segment is 42.8% (3 occurrences of word KRAH out of 7 words in total). On the other hand, the words are short, therefore length-based measures consider them to be simple words.

[5] The readability scores for the National Assembly subcorpus show high variability across the measures, which might be attributed to the fact that it is a different genre (spoken, but specialized). E.g., in several measures where the readability complexity rises from children's magazines to general magazines and newspapers, the National assembly scores are close to general magazines. Very long words are less likely used in spoken language, even in a political context. Average morphological difficulty and context of difficult words lead to the interpretation that this genre is more complex (less “readable”). The very high score for the context of difficult words might be attributed to enumeration of Assembly members (e.g., “Obveščen sem, da so zadržani in se današnje seje ne morejo udeležiti naslednje poslanke in poslanci: Ciril Pucko, Franc Kangler, Vincencij Demšar, Branko Kalalemina, ...” (I was informed that the following deputies are occupied and cannot attend this session: …). The relatively high percentage of verbs can also be interpreted from this perspective, e.g., the National assembly text include many performatives, such as “Pričenjam nadaljevanje seje” (Starting the continuation of the session) and “Ugotavljamo prisotnost v dvorani” (Establishing the presence).

[6] In summary, using a list of common words did not improve the partitioning of the text subcorpora perceived as easy and as difficult to read. Both measures that use it (Dale-Chall and Spache readability formulas) are poor separators. A number of simple readability measures worked well, such as the percentage of long words, the percentage of verbs/adjectives, and the average morphological difficulty.

[7] We also calculated the sample mean and standard deviation of readability measures for each text subcorpus. The results are shown in Table 2.

Table 2: The mean and standard deviation for each subcorpus of texts and each readability score.

Measure	Children's mag.	Magazines	Newspapers	Technical mag.	National assembly
% long words	0.065 (0.015)	0.109 (0.011)	0.137 (0.029)	0.146 (0.010)	0.137 (0.046)
Number of words	5488 (6184)	33966 (34821)	12881 (84708)	110875 (151007)	58841 (106515)
% adjectives	0.078 (0.016)	0.111 (0.013)	0.120 (0.020)	0.120 (0.008)	0.096 (0.022)
% verbs	0.216 (0.026)	0.170 (0.015)	0.161 (0.034)	0.144 (0.013)	0.180 (0.044)
% unique words	0.517 (0.077)	0.375 (0.053)	0.513 (0.114)	0.244 (0.144)	0.277 (0.173)
Context of difficult words	0.756 (0.054)	0.834 (0.027)	0.849 (0.133)	0.808 (0.036)	0.929 (0.044)
% difficult words	0.464 (0.048)	0.369 (0.022)	0.356 (0.122)	0.389 (0.032)	0.280 (0.036)
Gunning Fog Index	9.950 (1.255)	14.272 (1.271)	18.662 (9.319)	17.470 (0.800)	15.901 (3.493)
Flesch reading ease	37.592 (4.989)	23.855 (5.217)	10.002 (24.128)	12.520 (4.340)	19.178 (13.098)
Flesch–Kincaid grade level	10.500 (0.894)	13.596 (1.193)	17.356 (8.959)	15.999 (0.741)	14.523 (2.761)
Dale–Chall	2.845 (0.425)	4.036 (0.306)	4.972 (1.270)	4.941 (0.258)	4.560 (0.971)
Dale–Chall with word list	7.781 (0.720)	6.534 (0.357)	6.643 (2.163)	6.955 (0.484)	5.208 (0.539)

Spache readability formula	6.217 (0.368)	6.079 (0.348)	6.977 (3.499)	6.685 (0.323)	5.482 (0.600)
Automated readability index	12.873 (1.086)	16.117 (1.428)	20.474 (11.456)	19.007 (0.885)	17.014 (3.371)
SMOG	12.206 (0.759)	15.095 (1.066)	18.200 (2.757)	17.194 (0.611)	15.849 (2.500)
LIX	33.676 (3.384)	44.999 (3.282)	56.016 (23.123)	53.260 (2.077)	47.909 (9.073)
RIX	2.381 (0.496)	4.481 (0.781)	7.370 (3.836)	6.354 (0.518)	5.250 (2.574)
Coleman-Liau index	17.785 (1.120)	19.823 (0.861)	21.220 (1.807)	21.762 (0.903)	20.318 (2.170)
Avg. morphological difficulty	0.419 (0.017)	0.428 (0.010)	0.436 (0.044)	0.441 (0.017)	0.445 (0.026)
Avg. sentence length	8.353 (0.820)	13.389 (2.843)	21.120 (4.043)	18.641 (1.960)	19.063 (3.826)

[8] Using these results, we calculated the Bhattacharyya distance between the distributions of Children's magazines and newspapers for each score. The Bhattacharyya distance measures the similarity between two statistical distributions. We assumed the scores were distributed normally, as the results shown in Figure 1 show that the scores approximately follow a normal distribution, and calculated the distance using the following formula:

$$D_B(p,q)= \frac{1}{4}\ln\left[\frac{1}{4}\left(\frac{\sigma_p^2}{\sigma_q^2}+\frac{\sigma_q^2}{\sigma_p^2}+2\right)\right]+\frac{1}{4}\left(\frac{\left(\mu_p-\mu_q\right)^2}{\sigma_p^2+\sigma_q^2}\right)$$

[9] We also show the Bhattacharyya coefficient, which measures the overlap between two statistical distributions and can be calculated as:

$$\mathrm{BC}(p,q)= e^{(-D_B(p,q))}$$

[10] The results are presented in Table 3. These results are similar to the ones shown in Figure 2, with the readability formulas using the list of difficult words showing less dichotomization power. The largest distance is obtained using average sentence lengths.

Table 3: The Bhattacharyya distances and coefficients between the distributions of scores for children's magazines and newspapers for each readability measure. The results are sorted by decreasing distance.

Measure	**Distance**	**Coefficient**
Average sentence length	**2.866**	**0.057**
SMOG	1.433	0.239
% long words	1.350	0.259
RIX	1.101	0.333
Flesch-Kincaid grade level	0.956	0.385
Automated readability index	0.945	0.389
Dale-Chall readability formula	0.885	0.413
Gunning fog index	0.880	0.415
LIX	0.853	0.426
Spache readability formula	0.797	0.451
Flesch reading ease	0.776	0.460
% adjectives	0.719	0.487
Coleman-Liau index	0.708	0.493
% verbs	0.432	0.649
% difficult words	0.365	0.694
Dale-Chall with word list	0.318	0.728
Context of difficult words	0.285	0.752
Avg. morphological difficulty	0.235	0.790
% unique words	0.039	0.961

8. Additional Statistical Tests

[1] In addition to the initial analysis presented in the previous section, we performed additional, more

thorough statistical tests to determine which of the evaluated measures are better at predicting the group a text belongs to. We used the following approaches:

- **Mutual information.** This measure reports the amount of information we get about a random variable *Y* by observing another random variable *X*. In our case, mutual information reports the amount of information we get about the group of texts by knowing a score of certain readability measure. Mutual information is defined as:

$$\sum_{y\in Y}\sum_{x\in X} p(x,y)log\left(\frac{p(x,y)}{p(x)p(y)}\right),$$

  where p(x) and p(y) are the marginal probability distribution functions of *X* and *Y* and p(x, y) is the joint probability function of *X* and *Y.* In our case, X represents the distributions of readability measures and Y the distribution of groups. The higher the mutual information between the readability measure and the groups, the more useful the measure for determining the group membership.
- **Analysis of variance (ANOVA).** This measure first splits samples of a statistical distribution into several groups (in our case, based on the group the texts belong to) and then calculates if the groups are significantly different from one another. We use this measure to determine if the distributions obtained by calculating a single measure on each group of texts are significantly different. If they are, they can be useful for determining the group membership of a given text.
- **Feature selection using a chi-squared test.** Similarly to mutual information, we use the chi-squared test to determine whether the readability measures and the group memberships are mutually dependent. If they are, this indicates that knowing the value of the readability measure is useful when determining which group a text belongs to.

[2] In addition to the four statistical tests used above, we also ranked each feature using a random forest classifier (Breiman 2001). The classifier is capable of automatically combining different readability measures in order to predict which subcorpus a given text belongs to and is also capable of calculating how important each readability measure was when making the prediction. The classifier is described in more detail in the next section. Using each of these tests, we obtained scores that tell us how useful each readability measure is when trying to predict the subcorpus it came from. The results are presented in Table 4, with higher scores indicating better (more informative) readability measures.

Table 4: The ranks of readability measures obtained by the statistical tests, which report the usefulness of readability measures for predicting group membership. The measures are ordered from the most useful to the least useful.

Random Forest	ANOVA	Mutual information	Chi2

Average sentence length	Average sentence length	Average sentence length	% new words
% new words	% difficult words SPG	RIX	Number of words
Number of words	% long words	SMOG	% unique words
% unique words	SMOG	Percentage of new words	Flesch reading ease
% difficult words SPG	Dale-Chall	Automated readability index	LIX
Gunning fog index	Percentage of adjectives	Gunning fog index	Average sentence length
Percentage of verbs	Coleman-Liau index	LIX	% difficult words
RIX	Percentage of unique words	Number of words	Gunning fog index
Dale-Chall (word list)	RIX	Flesch-Kincaid grade level	Automated readability index
SMOG	% verbs	Flesch reading ease	% difficult words SPG
LIX	Flesch reading ease	Dale-Chall	Flesch-Kincaid grade level
Flesch-Kincaid grade level	Context of difficult words	% unique words	SMOG
Context of difficult words	LIX	% long words	RIX
Dale-Chall	Gunning fog index	% difficult words	Coleman-Liau index
% long words	Flesch-Kincaid grade level	% difficult words SPG	Dale-Chall
% difficult words	% difficult words	Spache readability formula	Spache readability formula
Avg morphological difficulty	Automated readability index	Context of difficult words	Dale-Chall (word list)

Automated readability index	% new words	Coleman-Liau index	% long words
% adjectives	Number of words	% verbs	Context of difficult words
Flesch reading ease	Dale-Chall (word list)	% adjectives	% verbs
Spache readability formula	Spache readability formula	Dale-Chall (word list)	% adjectives
Coleman-Liau index	Avg morphological difficulty	Avg morphological difficulty	Avg morphological difficulty

[3] The results of the statistical tests show that the features commonly used by the readability formulas (i.e. an average sentence length and number of long words) are useful when it comes to determining group membership. In particular, the average sentence length stands out since it is ranked as the most important measure in three out of the four tests. At least one of either LIX or RIX is also highly ranked (in the top 50% of all measures) by all the tests. Those measures are the only ones from the tested measures that were not designed specifically for English, which could be one of the reasons why they perform better on Slovene texts. The results also show that a number of proposed simpler readability criteria, such as the percentage of verbs, percentage of adjectives, and the average morphological difficulty are less useful than the established statistical formulas. The results are inconclusive about the most useful readability criterion for Slovene. Several formulas and statistics are useful, but the rankings are different by different tests. When using our list of common words Dale-Chall and Spache readability formulas are again shown to perform worse than the formulas that consider long words as difficult.

9. Classification Results

[1] In addition to statistical evaluation, we also performed a test with machine learning classifiers (Kononenko and Kukar 2007) to see whether we could use our readability measures to predict which subcorpus a text belongs to. With classification models, we can automatically learn how to split the texts into different subcorpora based on readability formulas and other readability criteria. We used the following classification models.

- **Decision trees** construct a binary decision tree where each node splits the training set based on one readability measure. The trained tree can predict the subcorpus of a given text.
- **Random forests (Breiman 2001)** create multiple decision trees in a random manner. This reduces the variance of a model and often gives better prediction accuracy than using a single decision tree.
- **Naive Bayes** is a probabilistic model based on the Bayes’ theorem. The model assumes that the readability measures are independent.
- **Extreme gradient boosting (Chen and Carlos 2016)** constructs a large number of simple

classifiers and combines them to achieve state-of-the-art results on many classification problems.

[2] In order to use classification models, we first train them on a training subset of our data set. We used randomly selected 75% of our data set for the training. To evaluate the models, we calculated the classification accuracy (i.e. the percentage of texts each model predicted correctly) on the remaining 25% of the data set. The obtained results are presented in Table 5. The results obtained by the majority classifier (i.e. classifying everything as the most frequent group) are presented as a baseline score.

Table 5: The classification accuracies for each of the models. The numbers show the percentage of texts for which the group membership was correctly predicted.

Model	Classification Accuracy
Random Forest	**0.984**
Extreme Gradient Boosting	0.979
Decision Tree	0.960
Majority Classifier	0.791
Naive Bayes	0.553

[3] Table 5 shows that we are able to predict the correct group of a text with high accuracy, over 98% with the best-performing model (Random forest). This shows that a combination of readability measures that we evaluated in this paper can be used to accurately distinguish between different groups of text.

10. Conclusion and Future Work

[1] We analyzed statistical distributions of well-known readability measures on Slovene texts. We extracted five subcorpora of texts from the Gigafida corpus with commonly perceived different readability levels: children magazines, popular magazines, newspapers, technical magazines, and national assembly texts. We find that the readability formulas are able to distinguish between these subcorpora reasonably well, with the exception of national assembly texts, which are of a different, spoken, genre and the used measures were not originally designed to handle it. A number of simple readability statistics, such as the context of difficult words and average sentence length, also dichotomize the different subcorpora of text.

[2] In this work, we only focused on simple readability formulas along with some additional readability criteria. There exist several more complex methods for evaluating the complexity of texts, such as the one presented in Lu (2009) and Wiersma et al. (2010). Such advanced methods might be more suitable for Slovene texts than the simple methods used in this paper, and we plan to test them in future work.

[3] Most of the used English readability formulas were designed to correlate with school grades and were initially tuned on that domain. For Slovene, there currently is no publicly available data set with texts tagged

according to the appropriate grade level. This disallows analysis of the readability measures from this perspective. In future work, we plan to prepare such a corpus and design several readability scores fit for different purposes. This will allow us to frame text complexity as a classification problem with the goal of predicting the grade level of a text instead of predicting its group membership. In a similar approach, experts would annotate texts with readability scores. This would allow us to fit a regression model using the readability measures analyzed in this paper.

[4] Another area that we plan to explore is the use of coherence and cohesion measures (Barzilay and Lapata 2008; Crossley et al. 2016), which are used to determine if words, sentences, and paragraphs are logically connected. Coherence and cohesion methods usually use machine learning approaches that mostly rely on language-specific features and shall be therefore evaluated on Slovene texts. The same applies to readability measures based on machine learning (Francois and Miltsakaki 2012) which we also plan to analyze in the future.

11. Acknowledgments

[1] The research was financially supported by the Slovenian Research Agency through project J6-8256 (New grammar of contemporary standard Slovene: sources and methods), project J5-7387 (Influence of formal and informal corporate communications on capital markets), a young researcher grant, research core fundings no. P6-0411 and P2-0103; Republic of Slovenia, Ministry of Education, Science and Sport/European social fund/European fund for regional development/European cohesion fund (project Quality of Slovene textbooks, KaUč). This work has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 825153 (EMBEDDIA).

Sources and literature

Literature:

- Anderson, Jonathan. 1983. “LIX and RIX: Variations on a little-known readability index.” *Journal of Reading* 26, No. 6: 490-96.
- Arhar Holdt, Špela. 2009. “Učni korpus SSJ in leksikon besednih oblik za slovenščino.” *Jezik in slovstvo* 54, No. 3-4: 43-56.
- Bailin, Alan, and Ann Grafstein. 2016. *Readability: Text and context*. Springer.
- Barzilay, Regina, and Mirella Lapata. 2008. “Modeling local coherence: An entity-based approach.” *Computational Linguistics* 34, No. 1: 1-34.
- Björnsson, Carl Hugo. 1968. *Läsbarhet*. Liber.
- Breiman, Leo. 2001. “Random forests.” *Machine learning* 45, No. 1: 5-32.
- Chen, Tianqi, and Carlos Guestrin. 2016. “Xgboost: A scalable tree boosting system.” In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 785-794. ACM.
- Coleman, Meri, and Ta Lin Liau. 1975. “A computer readability formula designed for machine scoring.” *Journal of Applied Psychology* 60, No. 2: 283.
- Crossley, Scott A., Kristopher Kyle, and Danielle S. McNamara. 2016. “The tool for the automatic

analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion." *Behavior research methods* 48, No. 4: 1227-37.

- Dale, Edgar, and Jeanne S. Chall. 1948. "A formula for predicting readability: Instructions." *Educational research bulletin*: 37-54.
- Dębowski, Łukasz, Bartosz Broda, Bartłomiej Nitoń, and Edyta Charzyńska. 2015. "Jasnopis–A Program to Compute Readability of Texts in Polish Based on Psycholinguistic Research." In *Natural Language Processing and Cognitive Science*, edited by B. Sharp, W Lubaszewski and R. Delmonte, 51-61. Liberia Editrice Cafoscarina.
- Fišer, Darja, Tomaž Erjavec, Ana Zwitter Vitez, and Nikola Ljubešić. 2014. "JANES se predstavi: metode, orodja in viri za nestandardno pisno spletno slovenščino." In *Language technologies : proceedings of the 17th International Multiconference Information Society - IS 2014*, edited by Tomaž Erjavec and Jerneja Žganec Gros, 56-61. Ljubljana: Jožef Stefan Institute.
- François, Thomas, and Eleni Miltsakaki. 2012. "Do NLP and machine learning improve traditional readability formulas?" In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, edited by Sandra Williams, Advaith Siddharthan and Ani Nenkova, 49-57. Association for Computational Linguistics.
- Grčar, Miha, Simon Krek, and Kaja Dobrovoljc. 2012. "Obeliks: statisticni oblikoskladenjski oznacevalnik in lematizator za slovenski jezik." In *Proceedings of the Eighth Language Technologies Conference,* edited by Tomaž Erjavec and Jerneja Žganec Gros, 89-94. Ljubljana: Jožef Stefan Institute.
- Gunning, Robert. 1952. *The technique of clear writing*. McGraw-Hill.
- Justin, J. 2003. *Učbenik kot dejavnik uspešnosti kurikularne prenove: poročilo o rezultatih evalvacijske študije.*
- Kilgarriff, Adam, Frieda Charalabopoulou, Maria Gavrilidou, Janne Bondi Johannessen, Saussan Khalil, Sofie Johansson Kokkinakis, Robert Lew, Serge Sharoff, Ravikiran Vadlapudi, and Elena Volodina. 2014. "Corpus-based vocabulary lists for language learners for nine languages." *Language resources and evaluation* 48, No. 1: 121-63.
- Kincaid, J. Peter, Robert P. Fishburne Jr, Richard L. Rogers, and Brad S. Chissom. 1975. *Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease formula) for navy enlisted personnel*. Report No. 8-75.
- Kononenko, Igor, and Matjaž Kukar. 2007. *Machine learning and data mining*. Chichester, Horwood Publishing.
- Kosem, Iztok, Tadeja Rozman, and Mojca Stritar. 2011. "How do Slovenian primary and secondary school students write and what their teachers correct: A corpus of student writing." In *Proceedings of Corpus Linguistics Conference 2011, ICC Birmingham*, 20-22.
- Logar Berginc, Nataša, and Simon Šuster. 2009. "Gradnja novega korpusa slovenščine." *Jezik in slovstvo* 54: 57-68.
- Logar Berginc, Nataša, Miha Grčar, Marko Brakus, Tomaž Erjavec, Špela Arhar Holdt, Simon Krek, and Iztok Kosem. 2012. *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES:*

gradnja, vsebina, uporaba. Ljubljana: Trojina, zavod za uporabno slovenistiko and Faculty of Social Sciences.

- Lu, Xiaofei. 2009. "Automatic measurement of syntactic complexity in child language acquisition." *International Journal of Corpus Linguistics* 14, No. 1: 3-28.
- Mc Laughlin, G. Harry. 1969. "SMOG grading - a new readability formula." *Journal of reading* 12, No. 8: 639-46.
- Senter, R. J., and Edgar A. Smith. 1967. *Automated readability index*. Ohio; University of Cincinnati.
- Sherman, Lucius Adelno. 1893. *Analytics of literature: A manual for the objective study of English prose and poetry*. Boston: Ginn.
- Škvorc, Tadej, Simon Krek, Senja Pollak, Špela Arhar Holdt, and Marko Robnik-Šikonja. 2018. "Evaluation of Statistical Readability Measures on Slovene texts." In *Proceedings of the conference on Language Technologies & Digital Humanities 2018,* edited by Darja Fišer and Andrej Pančur, 240-47. Ljubljana: Ljubljana University Press, Faculty of Arts.
- Spache, George. 1953. "A new readability formula for primary-grade reading materials." *The Elementary School Journal* 53, No. 7: 410-13.
- Verdonik, Darinka, Ana Zwitter Vitez, and Hotimir Tivadar. 2011. *Slovenski govorni korpus Gos*. Trojina, zavod za uporabno slovenistiko.
- Wiersma, Wybo, John Nerbonne, and Timo Lauttamus. 2010. "Automatically extracting typical syntactic differences from corpora." *Literary and Linguistic Computing* 26, No. 1: 107-24.
- Zwitter Vitez, Ana. 2014. "Ugotavljanje avtorstva besedil: primer »Trenirkarjev«." In *zbornik Devete konference Jezikovne Tehnologije Informacijska družba – IS*, edited by Tomaž Erjavec and Jerneja Žganec Gros, 131-34. Ljubljana: Jožef Stefan Institute.

Tadej Škvorc, Simon Krek, Senja Pollak, Špela Arhar Holdt, Marko Robnik-Šikonja

PREDICTING SLOVENE TEXT COMPLEXITY USING READABILITY MEASURES

SUMMARY

[1] In English, the problem of determining text readability (i.e. how easy a text is to understand) has long been a topic of research, with its origins in the 19th century. Since then, many different methods and readability measures have been developed, often with the goal of determining whether a text is too difficult for its target age group. Even though the question of readability is complex from a linguistic standpoint, a large majority of existing measures are based on simple heuristics. Since most of these measures were developed for English texts, it is hard to say how well they would perform on Slovene texts. Measures designed for English are designed to correspond with the American school system, are sometimes based on pre-constructed lists of easy words which do not exist for Slovene and do not take into account morphological information when determining whether a word is difficult or not.

[2] In our work, we analyze some common readability measures on Slovene text. We also introduce and analyze two additional readability criteria that do not appear in any of the analyzed readability measures:

morphological difficulty, where we assume word forms that appear rarely are harder to understand than the ones that appear commonly and the **context of difficult words,** where we assume difficult words are easier to understand in a context of simple words, as their meaning can be inferred from that context. We performed the analysis on 14,581 text documents from the Gigafida corpus, which were split into five groups based on their target audience (childrens' magazines, pop magazines, newspaper articles, computer magazines, and transcriptions of sessions of the National Assembly). We assumed that the groups should have different readability scores due to their differing target audiences and writing styles.

[3] For each analyzed readability measure we checked how well it separates texts from different groups. We did this by first obtaining the statistical distribution of readability scores for texts in each group and checking how much the distributions differ. We show that a number of common readability measures designed for English work well on Slovene texts. To determine which of the measures perform the best we used several statistical tests.

[4] We also show that machine-learning methods can be used to accurately (over 98% chance of a correct prediction) predict which group a text belongs to based on its readability scores. We trained four different machine-learning models (decision trees, random forests, naïve Bayes classifier, and extreme gradient boosting) and evaluated them on our dataset. We obtained the best result (98.4% classification accuracy) by using random forests.

Tadej Škvorc, Simon Krek, Senja Pollak, Špela Arhar Holdt, Marko Robnik-Šikonja

NAPOVEDOVANJE KOMPLEKSNOSTI SLOVENSKIH BESEDIL Z UPORABO MER BERLJIVOSTI

POVZETEK

[1] Problem berljivosti (t.j. kako enostavno je besedilo za branje) je v angleščini dobro raziskan. Obstaja veliko različnih metod in formul, s katerimi lahko analiziramo angleška besedila z vidika berljivosti. Kljub temu, da je vprašanje berljivosti z lingvističnega vidika zapleteno večina metod za ugotavljanje berljivosti temelji na preprostih značilnostih besedil. Ker je bila večina mer berljivosti zasnovanih za angleška besedila, ne moremo biti prepričani da bodo enako dobro delovala na slovenskih besedilih. Angleške mere berljivosti so namreč usklajene z ameriškim šolskim sistemom, včasih temeljijo na vnaprej sestavljenih seznamih lahkih besed in ne upoštevajo težavnosti besed z morfološkega vidika.

[2] V našem delu analiziramo pogoste mere berljivosti na slovenskih besedilih. Poleg tega uvedemo in analiziramo dva dodatna kazalnika berljivosti ki ne nastopata v pogostih merah berljivosti: **morfološka zahtevnost besed**, s katero želimo zajeti predpostavko da so redkejše morfološke oblike besed težko berljive, in **kontekst težkih besed**, s katero želimo zajeti predpostavko, da so neznane besede, ki se pojavijo v kontekstu znanih besed lažje berljive, saj lahko njihov pomen razberemo iz konteksta. Analizo smo izvedli na 14,581 besedilih iz korpusa Gigafida, ki smo jih razdelili v pet skupin glede na njihovo ciljno publiko (Otroške revije, splošne revije, časopisni članki, računalniške revije in transkripcije sej Državnega

zbora). Predpostavili smo, da imajo revije zaradi različnih ciljnih publik in tematik različne sloge pisanja in posledično različne stopnje berljivosti.

[3] Za vsako izmed mer berljivosti smo preverili, kako dobro med seboj loči besedila iz različnih skupin. Za vsako izmed njih smo pridobili statistično distribucijo vrednosti berljivosti vsake skupine in preverili, ali so distribucije ustrezno ločene. V analizi pokažemo, da se številne uveljavljene mere, ki so bile zasnovane za angleščino, dobro obnesejo tudi na slovenskih besedilih. Da bi ugotovili, katere mere najbolje razlikujejo med skupinami smo uporabili statistične teste.

[4] Poleg tega pokažemo, da lahko z modeli strojnega učenja in kombinacijo analiziranih metod berljivosti z visoko točnostjo (nad 98%) napovemo, v katero skupino spada določeno besedilo. Za to analizo smo uporabili štiri različne metode strojnega učenja (odločitvena drevesa, naključne gozdove, naivni Bayesov klasifikator, in extreme gradient boosting). Najboljši rezultat (98,4%) smo dobili z metodo naključnih gozdov.

Notes:

[*] University of Ljubljana, Faculty of Computer and Information Science, Večna Pot 113, SI-1000 Ljubljana, Jožef Stefan Institute, Jamova cesta 39, SI-1000 Ljubljana, tadej.skvorc@fri.uni-lj.si (mailto:tadej.skvorc@fri.uni-lj.si)

[**] Jožef Stefan Institute, Jamova cesta 39, SI-1000 Ljubljana, University of Ljubljana, Faculty of Arts, Aškerčeva 2, SI-1000 Ljubljana, simon.krek@guest.arnes.si (mailto:simon.krek@guest.arnes.si)

[***] Jožef Stefan Institute, Jamova cesta 39, SI-1000 Ljubljana, senja.pollak@ijs.si (mailto:senja.pollak@ijs.si)

[****] University of Ljubljana, Faculty of Arts, Aškerčeva 2, SI-1000 Ljubljana, University of Ljubljana, Faculty of Computer and Information Science, Večna Pot 113, SI-1000 Ljubljana, spela.arharholdt@ff.uni-lj.si (mailto:spela.arharholdt@ff.uni-lj.si)

[*****] University of Ljubljana, Faculty of Computer and Information Science, Večna Pot 113, SI-1000 Ljubljana, marko.robnik@fri.uni-lj.si (mailto:marko.robnik@fri.uni-lj.si)