CHAPTER 1

Previous Research on Language Readability

1.1 Introduction

Martinc, Pollak, & Robnik-Šikonja (2021) state 'Readability is concerned with the relation between a given text and the cognitive load of a reader to comprehend it.' It thus depends on many factors: lexical, syntactic, cohesive and background knowledge. Some of these can be measured directly, particularly lexical features, but less so for syntax and cohesion, and background knowledge is very difficult to objectively measure. Thus, traditional readability measures consider lexical features such as word length, sentence length, word difficulty, n-grams (sequences of a number of consecutive words or characters), the type-token ratio (TTR) which is number of unique words in the text divided by the total number of words, and parts-of-speech (POS) related, such as number of nouns, or ratio of function words to content words. In this chapter we will follow the classification of Martinc et al. (2021) by first considering lexical features for use in readability measures, then go on to the next stage and look at syntactic and discourse properties of texts, even though they tend not to work as well as lexical or semantic properties of texts. Syntactic features include such things as average parse tree height, or the average length of a syntactic unit. Working with a French language corpus, Todirascu, François, Bernhard, Gala, & Ligozat (2016) studied sixty-five discourse features such as lexical chains (words in a text which are related by hyponymy or hypernymy) but found that they contributed little to the predictive power of text readability classifiers when compared with traditional formulas.

The third stage is machine learning, where readability is regarded as a classification (is a text of high or low readability), regression (how does readability vary as the value of a linguistic feature varies) or a ranking task – is one text more readable than another. Machine learning produces better prediction than traditional approaches, but holds a disadvantage in that we need a training set of texts each characterised by a group of linguistic features and a human-assigned category label. Related to this,

https://doi.org/10.1017/9781009067386.001 Published online by Cambridge University Press

another problem is that the feature sets might work well on one corpus of texts, but not another. Other approaches they consider include language modelling, and recent work on deep learning.

1.2 Traditional Readability Measures

Martinc et al. (2021) list a number of widely used traditional readability measures:

Gunning Fog Index (GFI) (Gunning, 1952):

$$GFI = 0.4\left(\frac{totalWords}{totalSentences} + 100\frac{longWords}{totalSentences}\right)$$

Long words are those which have more than seven characters. GFI is the number of years of formal schooling required to understand the text, so a higher GFI means lower readability.

Flesch Reading Ease (Kincaid, Fishburne, Rogers, & Chissom, 1975), where a higher value of FRE means the text is more readable:

$$FRE = 206.835 - 1.05\left(\frac{totalWords}{totalSentences}\right) - 84.6\left(\frac{totalSyllables}{totalWords}\right)$$

The Flesch-Kincaid Grade Level is an adaptation of the FRE to give the required grade level:

$$FKGL = 0.39\left(\frac{totalWords}{totalSentences}\right) + 11.8\left(\frac{totalSyllables}{totalWords}\right)$$

The Automated Readability Index (ARI) (Smith & Senter, 1967) also gives a value for the GFI.

$$ARI = 4.71\left(\frac{totalCharacters}{totalWords}\right) + 0.5\left(\frac{totalWords}{totalSentences}\right)$$

The Dale-Chall readability formula (DCRF) (1949) uses a list of 3,000 words that a typical fourth-grade student could easily understand. Other words are 'difficult.' If the list of 3,000 words is not available, we can assume that words with more than seven characters are difficult.

$$DCRF = 0.1579\left(\frac{difficultWords}{totalWords}\right) + 0.0496\left(\frac{totalWords}{totalSentences}\right)$$

https://doi.org/10.1017/9781009067386.001 Published online by Cambridge University Press

The Simple Measure of Gobbledygook (SMOG) measure, originally used for the health domain, gives the grade level (McLaughlin, 1969). Polysyllables are words with three or more syllables:

$$SMOG = 1.0430\sqrt{numberOfPolysyllables\frac{30}{TotalSentences}} + 3.1291$$

1.2.1 Derivation of Early Readability Formulas

A number of costly mistakes have been made involving US Navy personnel, due to the training materials which had to be read in order to do a job, being written at a level of difficultly much above the reading ability of the man assigned to do the task (Kincaid et al., 1975). Training manuals which were more difficult to read resulted in more errors. There is thus a significant need in the armed services for readability measures which ensure that the reading levels of training manuals are not greater than the reading ability (denoted by the grade level) of the people who need to learn from them. Reducing the reading difficulty of texts is less expensive and time consuming than offering remedial reading instruction to staff or selecting only those personnel with high reading ability. Most reading formulas have been developed using members of the general population but had hitherto not been developed for US Navy personnel in particular. It was felt that the high number of technical terms in the Navy training manuals would cause readability measures to return grade levels which were too high. This problem was originally addressed by Caylor, Swift, Fox & Ford (1972), who derived the FORCAST formula for the Army from readability experiments specifically with Army personnel, but this formula includes a count of the one-syllable words in a passage, which at that time (and even today) was difficult to automate.

The grade level (in the US education system) required to read each sample text when deriving a readability formula was assigned if 50 per cent of the subjects reading that text scored 35 per cent or better on a Cloze test. In this Cloze test, every fifth word was left blank, and the reader's task was to fill each blank in. Flesch (1948) estimated that a success rate of 35 per cent to 40 per cent on the Cloze test was equivalent to a score of 75 per cent in a multiple choice test, so this criterion could also be used to determine whether the text was readable by people at a certain grade level. The Cloze

test is preferable, as the questions are prepared by a more objective procedure. An example of the Cloze test is as follows:

> Preheating involves raising the temperature of the base metal or a section of the base metal above the ambient temperature before welding. (first line of welding instructions, underlined words omitted)

Once the passages were graded, the multiple regression technique described in Chapter 2 was used to derive a new formula specifically for Army use. Similarly, FORCAST, the Flesch Reading Ease formula and another readability measure called the ARI Fog Count can all be made specific to a given subset of the population by recalculating new regression formulas. The ARI test needs a modified typewriter, but the other two can be calculated without specialised equipment. Klare (1963) automated FORCAST and the Flesch measure with computer programmes which can count the number of syllables in a text, said to be accurate to ± 1 per cent. Since all three measures give their results in terms of the grade level required to understand the texts, they should in theory be interchangeable.

As long ago as 1934, Dale and Tyler wanted to determine the readability of health education articles collected from books, newspapers and magazines on the subject of health education. In doing so, they piloted much of the original methodology for deriving readability formulas for a given topic area and subset of the population, in this case adults with poor reading ability. They found twenty-nine characteristics of texts thought to influence readability and correlated these with comprehension using Pearson's product-moment regression coefficient. Ten of these correlated significantly with comprehension and three were retained for inclusion in a readability formula.

Gray and Leary (1935) found eighty-two factors related to readability by studying existing reading materials, and recommendations of adult students, teachers and academics. Eighteen of these factors would be difficult to use in readability formulas, such as 'image-bearing words' which 'defy objective measurement.' Others were discarded because they were rare in the sampled texts. Twenty factors were retained because they correlated significantly (with a correlation coefficient $r > .27$) with the comprehension test scores of adult readers.

Smith and Kincaid (1970) warn that readability measures are estimates of difficulty and should not be taken as indicators of writing quality. Deliberately trying to write according to the formula criteria, such as shortening sentence or word length, does not necessarily improve writing, and may even degrade it.

Preferred early measures are Dale-Chall for its predictive power, followed by the Flesch formula which has the advantage of not needing a special word list. Custom formulas or special adaptations of existing formulas are needed for special populations, such as children.

1.3 Machine Learning

There is a section on machine learning in general in Chapter 2. For now, let us describe it briefly as follows: Let us take a machine learning classifier as a black box, where we know what goes in and what comes out, but for now ignore the mechanism by which the input is transformed into the output. There are two stages to its operation. First, in the training phase, we submit examples of texts and with each, a readability score assigned by a human assessor. At the end of this phase, the machine learning classifier will have learned to distinguish between texts with different degrees of readability. The texts are characterised by a typically large set of linguistic features. Second, in the testing phase, new documents, also characterised by a set of their linguistic features, are shown to the classifier, which will now be able to automatically classify them.

Curtotti, McCreath, Bruce, Frug, Weibel, & Ceynowa (2015) were interested in the accessibility of legal language to e general public. Their corpus was built from the U.S. Code of Federal Regulations, which provided both training and testing data. Gold standard (definitive) labelling of the texts was achieved by crowdsourcing, where the texts are labelled online by a large number of paid volunteers. The texts were characterised by linguistic features extracted using the Stanford dependency grammar parser and the Stanford context free parser. This enabled them to predict the readability of legal texts with high accuracy.

Kanungo and Orr (2009) looked at the readability of web page summaries or 'gists', which are frequently given to accompany the 'hit lists' provided as output by commercial search engines. It would be time consuming and expensive to provide readability assessments manually, and this would not be possible to do in real-time when people were actually using the search engines. However, once the machine learning classifier had been trained, it was possible to assign a readability score to each web page gist automatically. Once again, the machine learning classifier outperformed traditional readability measures. The learning algorithm, called a 'gradient-boosted decision tree', enabled the later extraction of linguistic features thought to influence readability.

Kotani, Yoshimi, & Isahara (2011)'s work focused on the readability of English texts to Japanese learners of the language. The subjects were instructed to read each text carefully, until they were able to answer a comprehension question. So that they did not feel under pressure and rush the task, they were not told that the time taken to answer the questions was being recorded. They used regression to estimate text readability based on comprehension rate and lexical, syntactic and discourse factors, such as the number of pronouns, as this requires the discourse task of identifying their referents.

In a highly novel departure from characterising texts by their linguistic features, Victoria Yaneva (2016) collected eye-tracking data while subjects were reading web pages, and used the patterns in the gaze track to characterise the texts. She was interested in evaluating and improving web site access for people with an autistic spectrum disorder (ASD), and the collected eye movement traces were stored in the ASD corpus. For the training phase, sets of traces obtained from subjects reading texts preclassified as easy, medium or difficult were presented to the classifier. Then in the test phase, reading difficulty could be estimated from the gaze tracks alone. From these experiments, Yaneva was able to compile a set of guidelines for improving text readability.

Buse and Weimer (2008) studied the readability not of text, but of computer code. They expected beforehand that this would depend on such things as indentation, choice of identifier names and the presence or otherwise of natural language comments to explain the following code sections. They used the Weka machine learning toolbox, which enables the operation of many different machine learning classifiers. They found that the main features degrading code readability were the average number of identifiers per line, the average line length, the average number of brackets ('{' or '(') per line, and the maximum line length. A few features were found to improve the readability of computer code: the average number of blank lines, the average length of comments, and surprisingly, there was a small beneficial effect of average number per line of the arithmetic operation symbols '+', '*', '%', '/' and '-'.

1.3.1 Training Corpora for Machine Learning Approaches

An early example of a supervised learning approach (Schwarm & Ostendorf, 2005) used a Support Vector Machine (SVM) trained on a corpus built from the now defunct children's magazine *Weekly Reader*, which contains articles grouped according to four reader age classes. More

recent work using this corpus is described by Petersen and Ostendorf (2009). The OneStopEnglish corpus has three readability classes (Vajjala & Lučić, 2018). Vajjala and Lučić used 155 linguistic features and the Sequential Minimal Optimization (SMO) classifier. The WeeBit corpus has texts divided into five readability classes (see Vajjala & Meurers, 2012) who used an early neural network approach called a multilayer perceptron. The WeeBit corpus annotations correspond to the age groups 7–8, 8–9, 9–10, 10–14, and 14–16 years. The three younger groups consist of articles from Weekly Reader, while the texts for the older groups come from the BBC-Bitesize website. The original corpus was built by Vajjala and Meurers (2012), but a version was later cleaned and used by Xia, Kochmar, and Briscoe (2016). The final corpus consisted of 3,000 documents, with 600 per age category. It is not a parallel corpus, so the texts at each level are unrelated to each other.

The Newsela Corpus (Xu, Callison-Burch, & Napoles, 2015) texts come from the Newsela company that produces reading materials for primary and secondary school pupils. It is a parallel corpus (showing which sentences of one side correspond with which sentence or sentences of the other side) of original and simplified texts at each grade from the second to the eleventh. There is not always the same number of texts for each grade, but there are 10,786 texts overall, produced from 1,911 originals. Other corpora used for training readability classifiers are Britannica (with adult and elementary versions of the same texts), and Convolutional Neural Networks (CNN) (with adult and abridged versions of the texts). The Common European Framework of Reference for Languages (CEFR) corpus is small, but has five levels of graded texts for English learners. The OneStopEnglish corpus (Vajjala & Lučić, 2018) consists of aligned text at three reading levels, and was created for English as a Second Language (ESL) learners. The texts originally came from *The Guardian* newspaper, and were simplified in two stages. It consists of 189 texts each at three levels. Martinc et al. (2021) used some of these corpora, and also used a corpus of Slovenian school books.

1.4 Language Model Features

An example of a language model is a Markov model. In a Markov model, the likelihood of a word or Part of Speech (POS) being seen depends on its inherent likelihood and the identity of a number of previous symbols. To take an example at the character level, the likelihood of encountering the letter 'u' at a certain position in the text depends on how common the letter

https://doi.org/10.1017/9781009067386.001 Published online by Cambridge University Press

'u' is in general and the previous characters. If the immediately preceding character is 'q', then the next character is highly likely to be a 'u.' A language model trained on one corpus is evaluated according to how well it predicts a separate test sequence of words, using the measure of perplexity (PPL).

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

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 new document. Perplexity scores can be used as features in a classifier – see Xia et al., 2016.

One advantage of a language model trained on English is that it can easily be retrained (given a new probabilistic dictionary, which is a long list of words and their probabilities) to work in French as well as English (Collins-Thompson & Callan, 2004). While the Type-Token Ratio (TTR) performed better on commercially available articles with human readability ratings, this model worked best for web pages and very short texts of less than ten words. This was the result they hoped for, since they were interested in producing a search engine which returned web pages of a suitable readability level as well as being on a suitable topic.

The Flesch-Kincaid formula assumes that the text is at least 100 words long and uses clear-cut sentences, neither of which are guaranteed for web pages.

Traditional models use just two or three variables – statistical language models can use a very large number.

A number of previous readability measures use vocabulary lists, such as the Dale-Chall (1949) formula which uses a vocabulary of 3,000 words which are easy to understand. These word lists act as simple language models, since each word in the list is implicitly given a score of 1 (for present) and each word not in the list a score of 0 (for absent). The method of Collins-Thompson and Callan (2004) uses more sophisticated language models since each word in the model is given a probability score, which is the number of times it occurs in the training corpus divided by the total number of words in the training corpus. They are said to be 1-gram models, since they consider only the frequency of single words, not the likelihood of them occurring in a particular word sequence, which would require 2-gram or 3-gram models. In early experiments they found that the more abstract a word, the more likely it is to be

https://doi.org/10.1017/9781009067386.001 Published online by Cambridge University Press

used at higher readability levels. This opens up the possibility of distinguishing between possible grade levels of a text by using the relative frequencies of the words in them.

They built a corpus of 550 web pages which were specifically written for readers at various grade levels. In these texts, concrete words such as 'red' were found most often at low grade levels, most typically 2; more abstract words such as 'perimeter' found peak usage at grade level 6, falling off at either side; 'determine' was most used at grade level 11. These results show that distinguishing readability by grade according to the abstractness of the vocabulary is feasible. If $L(G_i|T)$ is the likelihood of text T having the readability of Grade i (where for example G_5 is Grade 5), C(w) is the count of word w in text T, and $P(w|G_i)$ is the probability value for word w in the training corpus for G_i, then

$$L(G_i|T) = \sum_{w \in T} C(w) log P(w|G_i)$$

Thus for each grade level we work out the likelihood of our text having that degree of readability (so we work out the formula twelve times, one for each grade level). To do this for one grade level, we find for every word the number of times it occurs in the text and multiply it by the logarithm (function on a standard calculator) of the word's probability value in the training corpus of the language model. These products are then all added together. The language model with the highest likelihood is chosen as the one corresponding to the most probable reading level of the text.

The language models are smoothed, so that words which are nonexistent or very rare in the training corpora have their probabilities increased by a small amount, so 'nothing is impossible.' Language models are often created from the Wikipedia dump.

1.5 Deep Learning

Machine Learning techniques like SVM tend to be highly accurate. A subfield of machine learning is called deep learning. Deep learning models have the additional advantage of not requiring prior feature extraction: the inputs tend to be raw text, and in the case of readability measurements, the outputs tend to be either reading level categories or real-valued readability scores. Deep learning has the disadvantages that they can take a long time to train, the programmer is responsible for setting the fine-tuning parameters, and they may be hard to debug. Deep learning stems from the older

https://doi.org/10.1017/9781009067386.001 Published online by Cambridge University Press

paradigm of neural networks, which are inspired by the biological network of the human brain. In a neural network, a set of nodes, which roughly represent brain cells, is arranged in layers, where the inputs (such as a set of linguistic features) are presented to the input layer, and the output (in our case) could be one node representing a readability value. Between the input layer and the output layer is the hidden layer of nodes. Each node has a cell body and a nerve called an axon, connecting it to other nodes. When a cell receives sufficient input from other nodes or the input features, it can 'fire' and stimulate nodes in the next layer. It is the number of layers which distinguish deep learning from the older neural networks. Deep Learning must have more than three layers, which has become possible relatively recently as modern computers have greater processing power than before. However, in this chapter, we use the terms deep learning and neural network interchangeably. Deep learning may be supervised (requiring a training set of examples previously categorised by human experts) or unsupervised (where it learns to categorise data without any human intervention). Commonly-used examples of deep learning, which we will cover briefly in this section are Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) and Long Short Term Memory (LSTM). Both CNN and LSTM have been widely and successfully used in the field of Natural Language Processing. It is thought that CNN is good at capturing local features of language, while LSTM is good at processing sequence data (like words in a sentence) and capturing long-distance dependencies (such as anaphora resolution) (Jian, Xiang, & Le, 2022). Neural models outperform n-gram language models because they look at a wider context. There is a general maximum of 5 for n-grams due to data sparseness (Martinc et al., 2021).

Readability measures are relevant to tasks like text simplification and for teachers selecting materials at the right level for their students (Nadeem & Ostendorf, 2018). Traditional methods using linguistic features as inputs have the advantage of being easy to interpret. However, they have the disadvantages of not working well with short texts, and not generalising well to new domains, such as from literature to science. Sheehan, Flor, & Naplitano (2013) showed that traditional readability measures tend to underestimate the reading level of literary texts, while they tend to over-estimate the reading level of 'informational' texts, such as those in history, science and mathematics. In fact, they recommend a two-stage process, with genre identification as a precursor of readability measurement. Nadeem and Ostendorf's (2018) neural approach succeeded in overcoming both problems. They used a type of neural network called a RNN which

requires no set of linguistic features as input, but instead is able to generalise a likely reading level from the raw text input.

Recurrent Neural Networks have been used in a number of text representation tasks, most notably the Word2Vec architecture of Mikolov, Karafiát, Burget, Černockỳ, & Khudanpur (2010), which represents a word in the input text by a vector (list) of the collocational strengths of other words found in its context. So-called attention mechanisms can be shown to improve text processing tasks such as classification and machine translation (Nadeem & Ostendorf, 2018). These enable the neural network to focus on those parts of the input text which most contribute to the decision on grade level.

Uses of readability measurement include book recommendation or classification, web page summarisation for use with search engines, and (the focus of this book) accessibility in the health domain (Azpiazu & Pera, 2019).

Although neural networks can work using only raw text for input, circumventing the task of linguistic feature extraction, the system described here (called Vec2Read) also accepts part-of-speech and morphological information, such as person and tense. Azpiazu and Pera (2019) present a multiattentive recurrent neural network architecture for multilingual readability assessment. A RNN is a development of a traditional neural network in which each layer accepts not only inputs from the previous layer, but also an input from the immediately adjacent node in the same layer, representing information from previous words in the text (as words in a text are dependent on each other). RNNs have the problem that they are difficult to train, but Azpiazu and Pera (2019) overcame this difficulty by using an extension of the RNN called a LSTM which enables its neurons to either remember or forget certain information. They also used other types of neural networks.

In recent years, neural networks have outperformed approaches based on linguistically motivated features. Deutsch, Jasbi, & Scheiber (2020) tried adding linguistic information to the output of a neural network, but found that this did not lead to significant improvement. They hypothesised that their deep learning neural networks had already learnt all the linguistic information that could be gleaned from the raw text, both surface level and deeper factors, and thus the addition of more linguistic information was superfluous. They combined the deep learning model outputs with the traditional linguistic features by using them all as features for SVMs, a traditional machine learning architecture. A similar approach is taken in Chapter 3 of this book, where outputs of traditional readability measures are used as inputs for an SVM, alongside raw linguistic features.

Yancey, Pintard, & François (2021) present the first readability model for French as a Foreign Language (FFL) based on a deep-learning model, which outperformed the state-of-the-art. As well as features generated by the outputs of deep learning, they used cognitive features. Both types of features were incorporated into an SVM. The cognitive features included the proportion of abstract and concrete words, where more concrete words tend to be more easily recognised. They also used the Orthographic Levenshtein Distance 20 (OLD20), where Levenshtein Distance is the number of edits (insertions, deletions or substitutions) required to transform one word into another. The OLD20 version returns the average Levenshtein distance taken over the closest twenty words found in the lexicon. The rationale is that the existence of a large number of orthographic neighbours (similarly spelled words) results in interference when reading, producing lower readability. The third cognitive criterion was the average number of commonly known senses per word, as found in the large semantic network Babelnet, which if greater, tends to produce poorer readability. Their evaluation criteria were accuracy, the proportion of times that the machine assigned texts the same grade level as human assessors, adjacent accuracy which is like accuracy but also allows the classifications to differ by 1, and the Pearson product-moment correlation coefficient between the set of humanly-assigned grade scores for each text and the corresponding machine-calculated grades.

1.6 Readability Measures in Other Languages

Azpiazu and Pera (2020) state that readability formulas are rarely consistent when measuring translations of the same text. There was thus a need to develop measures of readability for texts in languages other than English. The four stages in development of such measures seem to be firstly to use a formula designed for English exactly as it is, but on other language texts; secondly to adapt these English formulas according to the quantitative characteristics of the language pair, such as relative word length; thirdly to develop formulas completely afresh for the other language; and finally there is a trend in several languages to machine learning and deep learning approaches. This chapter will concentrate on three languages for which considerable work on readability has been done, namely Italian, Swedish and French. We will also briefly describe individual case studies for German and two Indian languages.

Chapter 5 will describe in detail some work on readability in Japanese. Sung, Chang, Lin, Hsieh, & Chang (2016) used an SVM with twenty-four

linguistic features to predict text readability in Chinese. Correia and Mendes (2021) describe a deep learning model for measuring the readability score of texts in European Portuguese. See Škvorc, Krek, Pollak, & Arhad Holdt (2018) for Slovenian.

1.6.1 Readability in Italian

The earliest readability index for Italian was Flesch-Vacca which was derived from the Flesch index for English. The Flesch index version they used was F = 206.385 – (84.6 × S) – (1.015 × P) where F is readability, S is the average number of syllables per paragraph, and P is the average number of words per sentence. Roberto Vacca and Valerio Franchina proposed the following adaptation for the Italian language:

$$F = 206 - (0.65 \times S) - P$$

Instead of counting the length of words and sentences in syllables, Gulpease (Lucisano & Piemontese, 1988) uses the number of letters, giving a simpler calculation, where counts for sentences, letters and words are those for the entire document.

$$F = 89 + \frac{300 \times numberofsentences - 10 \times numberofletters}{numberofwords}$$

F should be in the range 0 to 100, where 0 is the most difficult. However, it is not constrained to be in this range. For example, texts with F > 40 should be legible by students with high school diplomas, and texts with F > 60 can normally be read by students with a middle school diploma, and texts with F > 80 can normally be read by students with a primary school diploma.

The system of Tonelli, Tran Manh, & Pianta (2012) accepts documents in the Italian language, and outputs a set of readability metrics inspired by Coh-Metrix (Graesser et al., 2004; http://cohmetrix.memphis.edu), which is a package able to calculate about sixty readability-related aspects of an input text. Coh-Metrix takes advantage of recent work in psycholinguistics and natural language processing such as dependency and constituency parsers, and anaphora resolution. The system also displays how each measure compares with three levels of Italian education – elementary, middle and high school. The aim of the research by Tonelli et al. (2012) was to improve web content readability.

Three Coh-Metrix indices were derived by collecting texts written by three groups of people: children in elementary schools, children aged 11–13 in

https://doi.org/10.1017/9781009067386.001 Published online by Cambridge University Press

middle schools, and teenagers in high schools. A similar approach had been advocated by Crossley, Allan, & McNamara (2011), who used original and manually simplified texts. The linguistic features measured by Coh-Metrix are in five main classes: First, general word and text information such as syllables per word and the TTR. If these are high, it seems likely that the sentence would be more difficult to process cognitively and hence more difficult to read. Similarly sentences with greater syntactic complexity, such as those with a greater number of embedded constituents will be more difficult to read. Third, referential and semantic indexes. For example, if a sentence is not cohesive with respect to adjacent sentences it is more difficult to process. Fourth, indices for the complexity of a mental model evoked by a document (see Dijk & Kintsch, 1983) and lastly existing measures such as Flesch Reading Ease. The online demonstration of CohMetrix shows about sixty such measures. Syllables were computed using the Perl module Lingua::IT::Hyphenate: see http://search.cpan.org/~acalpini/Lingua-IT-Hyphenate.

The Italian adaptation of Coh-Metrix is called Coease. Here features 1 – 6 are basic features of the text such as number of words or syllables. Features 7 to 10 consider the familiarity of words using a frequency list from the whole of the Italian Wikipedia at that time – Low frequency words are less familiar. Index 8, the logarithm of the raw frequency of content words was used, as this psycholinguistic measure relates to reading time (Haberlandt & Graesser, 1985). Content words can be estimated as those with parts-of-speech verb, adjective and noun. Indices 11–12 measure the abstractness of nouns. One way of doing this would be to look at the level of the word in the MultiWordNet hierarchy (Pianta, Bentivogli, & Girardi, 2002). Indices 20–29 capture cohesion of sentences by looking at connectives. Indices 30–31 relate to the syntactic similarity of sentences – high syntactic variability in a text being easier to understand.

The purpose of assessing readability in the READ-IT project was to evaluate text simplification for newspapers. Potential beneficiaries were people with low literacy skills, both native speakers and learners of Italian. Health-related information was another crucial domain where it was desired to make the information available to a large and heterogeneous reader group. READ-IT used a SVM. Different linguistic features came into play according to whether the readers were L1 or L2 learners. Four levels of features proposed by the READ-IT project corresponded with the NLP tools required to calculate them.

First, the raw text features typically used in traditional metrics such as word and sentence length. The tools needed to extract them included

tokenisers. Second, they used lexical features such as those extracted by lemmatisers, including the lemmas in the Basic Italian Vocabulary by De Mauro (2000). This contains a list of 7,000 frequently used Italian words. They also used the TTR. Here there is a requirement for text samples of equal length, as TTR decreases as texts get larger and fewer new words are introduced. Morpho-syntactic features were extracted by such tools as part-of-speech (PoS) tagging, such as the Language Modelling probability of PoS unigrams, and lexical density, which is the ratio of content words to all tokens in the text. Verbal mood was a feature specifically for Italian. Syntactic tools were extracted by dependency parsers, which allowed the calculation of such measures as parse-tree depth features (including whole-tree parse tree depth), and the longest path from the root of the dependency tree to some leaf, and the length of dependency links (equal to the number of words occurring between the syntactic head and the dependent, or phrase length). The corpus used to extract the linguistic features was La Repubblica, a highly readable newspaper. The morpho-syntactic features were the most reliable, producing 98 per cent accuracy.

1.6.2 Readability in Swedish

The most widely used metric in Swedish is LIX, the Läsbarhets (Readability) Index of Björnsson (1968). Based on the Flesch score, it can be applied to texts written in any language, since it depends only on word length, sentence length and the number of 'difficult' words, namely those of more than six characters. LIX = 100(B/W) + (W/S) = per cent of long words plus average words per sentence. A score of 20 would suggest a simple text and one of 60 a highly complex text.

The OVIX (Ordsvariationsindex or Word Variation Index) and the Nominal Ratio (Hultman & Westman, 1977) were developed later as they correlated better with features taken from higher linguistic levels, which have been made possible to count by the advent of text analysis tools such as grammars and parsers. The OVIX formula is given by King (2019).

$$OVIX = \frac{\log(total\ \ number\ of\ \ words)}{\log\left(2 - \frac{\log(number\ of\ \ unique\ words)}{\log(total\ \ number\ of\ \ words)}\right)}$$

King gives these six easy steps for working out this formula:

1. Find the logarithm of the total number of words (tokens) in the sample.

https://doi.org/10.1017/9781009067386.001 Published online by Cambridge University Press

2. Find the logarithm of the number of unique words (types) in the sample.
3. Divide the result of step 2 by the result of step 1.
4. Subtract the result of step 3 from 2 (the number 2, not the result of step 2)
5. Find the natural logarithm of the result of step 4.
6. Divide the result of step 1 by the result of step 5.

Falkenjack, Heimann Mühlenbock, & Jönsson (2013) created a training corpus made up of simplified texts, called the LäsBarT corpus, and five corpora of non-simplified texts in the domains of general news, popular science, professional news, government text and fiction. This solved the problem of text annotation, as according to the origin of the texts they could automatically be labelled as 'easy-to-read' or 'not-easy-to-read.' The machine learning algorithm they used was SVM (used successfully in other readability experiments, such as Feng et al. (2010), who used the Waikato Environment for Knowledge Analysis (Weka) package (Garner, 1995)). As was done with the READ-IT system, a set of candidate features found in the literature was subdivided into shallow (requiring tokenisation), lexical (requiring lemmatisation), morpho-syntactic (needing part-of-speech tagging), and syntactic (requiring parsing). Best accuracy when using single feature models was obtained for unigram dependency type (the unigram probabilities for the sixty-three dependency types per token) closely followed by the unigram part-of-speech type per token, where each part-of-speech ratio acted as an individual attribute.

In this section we contrast traditional readability measures with an approach which combines automated linguistic analysis with machine learning. Falkenjack et al. (2013) follow a similar approach, and look at the extent to which linguistic features proposed for other languages, especially English, are suitable for determining readability in Swedish. They found that the best performing features were sequences of part-of-speech tags and syntactic dependency type.

Reasons for measuring readability in text include giving teachers a means of assessing a student's reading ability, and advising general readers that they might consider looking for easier texts on the same topic – and readability measures would help them find such texts. In this chapter we regard readability as meaning easy-to-read, although Dale and Chall (1949), who developed one of the earliest readability formulas, define readability also as a function of how well the readers understand the text, whether they can read it at a comfortable speed, and whether they find the text interesting.

Pilán et al. (2014) studied the readability of sentences understandable to second language learners of Swedish. Most previous studies have concentrated on L1 readers. Their rule-based approach slightly outperformed the combined machine learning rule-based model. It is useful to check the readability of sentences to make sure they are suitable for automatic grading exercises. To help select linguistic indicators, they used suggestions made in previous machine learning-based readability research, aspects of Good Dictionary Examples (GDEX) (Kilgariff et al., 2008) and second language learning syllabuses. Machine Learning was able to distinguish sentences suitable for intermediate learning ('within the B1 level') from those which could be used for advanced learners 'above the B1 level.'

Traditional readability measures for Swedish include the Läsbarhets index and the Nominal Ratio. Pilán et al.'s (2014) paper is concerned with sentence-level readability as opposed to full-text readability, as they were interested in selecting appropriate vocabulary example sentences. Traditional readability measures have not been found to work so well for very short texts (Kilgariff et al., 2008). GDEX is a sentence-ranking system, which credits sentences according to such things as length, the inclusion of simple vocabulary items and the lack of anaphoric pronouns. This study used the Korp corpus, where the words are mostly POS-tagged, syntactic dependency relations, lemmas and sense identifiers. They also used frequency-based word lists, one based on the Swedish Wikipedia, and the CEFR corpus, a collection of L2 Swedish course-book texts. All of the Korp corpus was B1, while CEFR consisted of some B1 and some B2 texts.

The features used were traditional readability measures, easy-to-compute shallow features such as sentence length including punctuation and mean number of characters per token. Syntactic features were used such as the ratio of a morphosyntactic category such as adverbs to the number of content words in a sentence. Lexico-morphological features such as TTR (vocabulary diversity) were also used, the rarity of words using the Wikipedia word list being found to work best. Semantic features, such as does a word have more than one meaning, as determined by the meaning (sense) codes in an automatic tagger? were used as well. It is possible to find the most important features 'weightings' after the event in machine learning – the best was percentage of words above the B1 level. Classification accuracy for the SVM was about 71 per cent, comparable to the value obtained by previous authors for similar tasks.

Larsson (2006) presented a SVM classifier for Swedish readability levels. The corpus consisted of quality newspaper texts as the more difficult-to-read, high-school texts to represent medium level readability, and easy-to-read tabloid newspapers.

1.6.3 *Readability in French*

A review of early readability measures for French is given by Benoit (1986). The earliest readability measures for French were adaptations of the American Flesch score. Kandel and Moles (1958) produced a light modification of the English language Flesch formula, taking into account that on average, words in French are a little bit longer than English words. More widely used is the adaptation of Landsheere (1963): Facilité = 206.835 – (X + Y). In general, there is a correlation between sentence or word length and difficulty. Thus X is a syntactic variable, namely the average length of sentences expressed in words. This number is multiplied by a coefficient of 1.015. Y is the lexical variable, the number of syllables per 100 words, multiplied by 0.85. This formula results in a value between 0 (practically illegible) and 100 (easy for any literate person). In deriving this formula, the corpus consisted of excerpts from textbooks with reading difficulty defined in the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) at six levels: A1 Breakthrough; A2 Waystage; B1 Threshold; B2 Vantage; C1 Effective Operational Proficiency and C2 Mastery (François and Fairon, 2012).

Landsheere's student, Georges Henry (1975), produced the first original set of formulas for French. The most complicated one has 26 parameters and requires computer calculation. The simplest, called the manual method, only three variables: syntax (number of words per phrase as in the Dale-Chall formula), and vocabulary not in the Gougenheim list of simple words (excluding proper nouns, and technical terms defined in the text), and dialogue and dramatisation markers such as first names, quotation marks and exclamation marks). Using two of these three variables, we may read off the Henry score from a set of curves on a graph. The curves on the graph appear one above another and correspond to the number of difficult words per hundred words (the greater the number, the lower the curve). We read off the syntactic variable on the horizontal axis, then look directly up to the relevant vocabulary curve. At the meeting point, we look directly left, and read off on the vertical axis the relevant readability score. This test was designed so that a text would ideally be in the range 35 to 45. Texts with scores over 45 would be too difficult, while those with scores less than 35 would be too easy. Three separate graphs were produced by Henry: one for first grade students, and two others for fourth and sixth grade students (Benoit, 1986).

François and Fairon (2012) used machine learning to distinguish between texts which learners of French as a foreign language (FFL) would find easy to read and which they would not. Readability features specific to FFL include

the fact that Multi-Word Expressions (MWE) are known to cause difficulty for students of a language. Also variables typical of dialogue are more common in lower grade books) e.g. punctuation types. To the best of our knowledge, Uitdenbogerd's (2005) formula for the readability of French as a foreign language is the first formula to incorporate cognates, which are 'words that are recognisably similar to words with the same meaning in the person's native language'. Some books intentionally make use of cognates such as Uitdenbogerd's own comic book which used only cognates and 12 of the 20 most common words in French newspapers. To estimate readability, she took the first 100 words of each text and then continued on to the next sentence break. Cognates were manually counted: either an exact spelling, plus or minus a terminal 'e', or a syllabic word with an 'obvious' common root such as 'compliqué' (complicated). If a cognate was found more than once, it was counted just once for each time it occurred. The readability FR = 10 × WpS – Cog, where WpS is the average words per sentence and Cog is the number of cognates. The work is interesting as this is a characteristic of French as a Foreign language – more basic textbooks contain more cognates. The problem is that it was developed for only one language pair, but the method would apply to any two languages which are historically related or contain a large number of borrowings.

Collins-Thompson and Callan (2004) describe a language modelling approach to predicting reading difficulty in French.

1.6.4 Readability in German

Vor der Brück and Hartrumpf produced a semantically oriented readability checker for German. They used a parser called WOCADI (Hartfumpf, 2003) which can transform a text into a semantic net, a structure where each discrete concept is described as a node, and the relations between them are described by links. The meaning of the text is captured by the use of labels on the nodes (such as the word 'today') and links (such as the relation 'temporal'). A list of characteristics of each node concept can be stored with the name of the concept. They found that a number of characteristics of the semantic net were related to readability. For example, a text is more readable if it contains more concrete and fewer abstract nouns; fewer negations; indicators concerning anaphors; number of prepositions in the sentence; and the longest path in the semantic net. This last feature is the greatest number of links which must be traversed to go from any node to any other one in the net.

1.6.5 *Readability in Indian Languages*

Pantula and Kuppusamy (2020) estimated grade level by measuring the readability of web pages in Indian languages. English readability formulas are not suitable for evaluating Hindi and Bangla texts because of morphology (languages with more complex morphology produce longer sentences) and different grapheme correspondences (different numbers of characters per sound). They collected over 8,000 unique web pages, then used the Flesch-Kincaid measure to classify these web pages in two ways: First as less readable, readable and highly readable, and secondly according to minimum grade level required to understand the texts using the grade level version of the Flesch-Kincaid formula. The linguistic features of the texts, extracted using packages available with the Python programming language, included average number of words per sentence, number of sentences, and the average number of syllables per sentence. This data was used to train the classifier. More specifically, they considered 6 'readability characteristic' features (word count, average number of words per sentence, number of sentences, syllable count, average number of syllables per sentence, and number of words containing more than two syllables). Many words in a number of languages have a CVC structure, where one consonant is followed by one vowel, which in turn is followed by a consonant. If a word has a CVCVC structure, then the number of vowels is the number of syllables. This simple method, however, would not work for words such as 'make' in English. The lexical features they used included type-token ratio, rare words, defined by words containing the rare (in English) characters b, j, k, q, v, w, x, and z and lexical ambiguity. The best performing machine learning models were called Random Forest (RF) and Decision Tree Classification (DTC). They developed a browser extension to assess every uploaded web page for readability.

1.7 Conclusion

Readability formulas are important in education as teachers need to 'match learners with texts' (François & Fairon, 2012). The first step in enhancing anything is to measure it (Pantula & Kuppusamy, 2022). Readability assessment has been a key research area for the past eighty years, and still attracts researchers today. The text is not the sole determiner of readability – it also depends on the audience, where for example a text might be readable to a native speaker but not to a learner. The most common measures currently in use are Flesch-Kincaid and Dale-Chall. In the Flesch-Kincaid measure,

number of syllables per word is a proxy for lexical complexity, while sentence length is a proxy for syntactic complexity (Lucisano & Piemontese, 1988).

Traditional models were parsimonious, incorporating as few linguistic features as possible, and used linear regression to combine two or three surface features. Later models used psychological theory, measuring such things as coherence, density, and inference load. The inference load for a text differs according to the user's skill and knowledge, and depends on recovering the event chains (actions, physical states and mental states) underlying the texts (Kemper, 1983). Petersen and Ostendorf looked at the inadequacy of reading formulae to argue for machine learning approaches. Thus, the period from 2000 to 2010 saw work on readability making use of more sophisticated NLP techniques such as syntactic parsing, statistical language modelling, and machine learning. The later chapters in this book look at non-traditional uses of readability measures, such as determining the suitability of texts for machine translation. In this vein, Zwitter Vitez (2014) used readability measures as features in an authorship attribution task. Sheehan et al., 2013, look at the applicability of readability formulas across genres. Another theme of this book is the focus on one genre in particular: texts on environmental health.

Liu, Ji, Shanshan Lin, Zhao, & Lyv (2021) state that readability measures based solely on linguistic features do not take account of the reader's age, health, cultural or linguistic background. Their paper describes the use of machine learning to combine linguistic features with reader background, enabling a move away from 'one size fits all' metrics. For example, words in lists such as the Dale-Chall list which form part of some readability measures may be easier to understand by some people than others. Factors include familiarity with the health education traditions in the language of the health-related texts. Difficult-to-read health information can result in confusion and medical errors.

In the study by Liu et al. (2021), texts on the topic of infectious diseases from high-quality Australian websites were rated by individual human readers, and these individual human assessments formed the gold standard for training and testing. The features were linguistic – hence we have a combination of linguistic and human readability data. The subjects were all learners of English with no specialist medical knowledge. A variety of machine learning models were used and one neural network. Key surface linguistic features were average syllables per word (showing morphological complexity) and sentence length, both form part of Flesch and Linsear, where the Linsear Write Formula is Score = (Easy Words) + (Hard Words x 3) / Sentence Count. If the number of easy words is less than twenty, divide

by 2, otherwise subtract 2 then divide by 2. The Machine Learning methods performed well. Various researchers have come to the conclusion that Machine Learning methods can improve readability estimation. The process is data-driven, requiring less manual labour and avoiding human bias (Liu et al., 2021). Current research seems to focus on deep learning methods, which show great promise.