CAS Report
Comparison of rule compliance in corpora for the Austrian and Brazilian Easy Languages

CAS Translation Technology and Artificial Intelligence
March 2022
by Christina Maria Müller

Supervisor: Prof. Dr. Silvia Hansen-Schirra
Universität Mainz

University of Zurich^UZH

Abstract

This assignment aims at analyzing the compliance of the Austrian and Brazilian Easy Languages with the German set of rules for Easy language. Corpora for both Austrian German and Brazilian Portuguese were created by extracting contents written in Easy Language from Austrian and Brazilian websites and online documents. Both corpora were then tested on five essential German Easy Language rules: average sentence length, separation of compounds, avoiding third-person personal pronouns, avoiding subordinate conjunctions and the employment of bulleted sentences. Both languages proved to apply the rules of short sentences and clear structuring through the use of bulleted sentences. With regard to the separation of compounds, it was established that in the Austrian corpus the use of hyphens in compounds is inconsistent, and that in Brazilian Easy Language this rule is not (yet) a concern. The small number of personal pronouns used showed major compliance with the set of rules applied for both languages whereas replacing subordinate conjunctions is still not a rule very consistently employed, which calls for stricter compliance with rules by the creators of Easy Language.

Keywords*:* Easy Language, Austria, Brazil, German Easy language rules, compliance

Contents

1 Introduction

The use of Easy Language has become more and more relevant in many countries, such as the European Union, Switzerland, Great Britain, the US, Australia, and Brazil, since 1970 and has been developing in terms of rules and application ever since. The first set of rules for Easy Language, “Make it simple : European guidelines for the production of easy-to-read information for people with learning disability for authors, editors, information providers, translators and other interested persons” (FREYHOFF, 1998), was published by the International League of Societies for Persons with Mental Handicaps over twenty years ago, in June 1998. It has become apparent that vast proportions of the people living in any country are not capable of receiving complexly written contents, be it due to cognitive disabilities, dyslexia, dementia, or simply because an individual’s mother tongue is not the one spoken in the country where he or she lives. Among several difficulties that this population is facing, it has also emerged the awareness several decades ago that writing in complex language in fact induces a violation of the civil rights of a human being (MENDONÇA, 1987) especially when it comes to information provided by the public sectors. As shown below, the number of the affected population speaks for itself.

Therefore, the objective of this assignment is to analyze and compare the practical application of five of the fundamental Easy Language rules in Easy Language texts from Austria and Brazil, based on the German set of rules for Easy Language published by Duden (BREDEL & MAAß, 2016), in order to understand if and to what extent the collected texts from Austria and Brazil comply with the analyzed rules of Easy Language.

1.1 Easy Language in Austria

According to the PIAAC study (OECD, 2016), almost a million Austrian inhabitants are considered functional illiterates. This amounts to 17.1% of inhabitants with considerable difficulties to receive texts written beyond the lower language proficiency levels (see below), which is almost always the case in contents provided by public authorities, be it from different ministries, the justice, or the healthcare system.

Especially after the adoption of the Austrian Act on the Elimination of Discrimination against People with Disabilities (Behindertengleichstellungsgesetz, BGG) on January 1st, 2006, the Austrian public sector has been striving to provide its contents in Easy Language.

Different from Germany where there is a distinction between Easy Language (Leichte Sprache), Easy Language Plus (Leichte Sprache Plus) and Plain Language (Einfache Sprache) in ascending level of difficulty, in Austria the classification of the different levels of Easy Language follows the European Reference Framework for Languages with levels A1, A2 and B1.

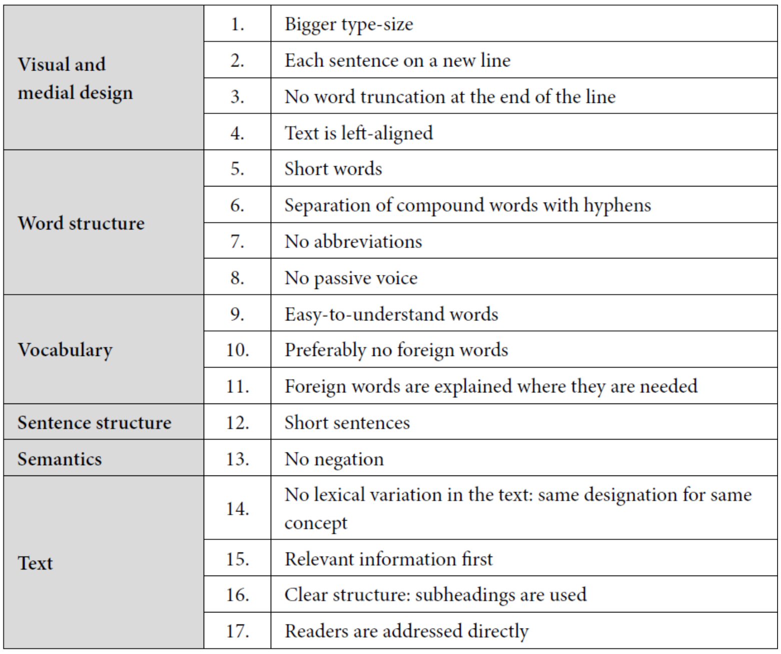
In Austria, the private company atempo Betriebsgesellschaft mbH founded the Capito brand in 2016 addressing three of the 17 UN Sustainable Development Goals namely no. 4 (quality education), no. 10 (reduced inequalities) and no. 16 (peace, justice and strong institutions). Capito elaborated a 150-item criteria catalog on Easy Language which it discloses to franchise partners only. Most Austrian public sector websites are translated into Easy Language by Capito as it can be seen by the specific Easy Language A1, A2, and B1 labels which are exclusively used in translations done by Capito itself or its franchise partners.

1.2 Easy Language in Brazil

In Brazil, the use of “linguagem simples” (easy language) or, previously, “linguagem clara” (plain language) (cf. (BARBOZA, 2010)) has also been discussed for decades particularly in the social and linguistic areas. According to a survey carried out by the non-governmental organization Ação Educativa (AÇÃO EDUCATIVA; INSTITUTO PAULO MONTENEGRO, 2018), no less than three out of ten Brazilians are considered functionally illiterate. Besides the employment of Easy Language as a means to combat social inequality, the advent of the coronavirus pandemic in 2019 gave further momentum to the topic as the public authorities became aware that immediate action was needed in order to keep also the most vulnerable portion of the population up to date regarding the pandemic. In response to that, the then councilman of the São Paulo Municipal Chamber, David Annenberg, elaborated a bill of law as the basis for the Municipal Policy of Easy Language (Política Municipal de Linguagem Simples). The corresponding Municipal Law no. 17,316 was adopted by mayor Bruno Covas on March 6th, 2020, applicable to the direct and indirect administration bodies of the

Municipality of São Paulo, including the Municipal Chamber and the Municipal Audit Court. While several other Brazilian states are working on similar official regulations, the Federal legislation includes the Brazilian Federal Law no. 12,527 from 2011 (on access to information provided in a transparent, clear and easy-to-understand language) and no. 13,146 from 2015 (on the inclusion of people with disabilities) as well as the bill of law PL 6256/2019 that establishes the National Policy of Easy Language in the bodies and entities of the direct and indirect public administration, which is yet pending enactment.

A specific Brazilian set of rules as such does not exist. However, there exist several guidelines elaborated by entities and initiatives in different states, such as the Brazilian Network for Easy Language (Rede Linguagem Simples Brasil), the innovation laboratories (011).lab of the city of São Paulo and Iris Lab Gov of the state of Ceará, the initiatives Comunica Simples and Movimento Down in the state of Rio de Janeiro, or the Easy Language and Justice Innovation Program (Programa Linguagem Simples & Inovação Jurídica) of the Audit Court of the state of Santa Catarina. These different guidelines cover mostly the same or similar items, and are also contained in the rules for the German language (no negation, no passive voice, no abbreviations, preferably no foreign words, etc.). According to Maaß et al., when combining the three most common sets of rules in Germany, the number of rules sums up to 120, which are necessary for translation and creation purposes. 17 out of the 120 rules coincide in all three guidelines, and they suffice to identify an Easy Language text as such. These 17 rules correspond largely to the ones found in the guidelines of the above-mentioned different institutions in Brazil, and are the following:



Visual and medial design	1.	Bigger type-size
	2.	Each sentence on a new line
	3.	No word truncation at the end of the line
	4.	Text is left-aligned
Word structure	5.	Short words
	6.	Separation of compound words with hyphens
	7.	No abbreviations
	8.	No passive voice
Vocabulary	9.	Easy-to-understand words
	10.	Preferably no foreign words
	11.	Foreign words are explained where they are needed
Sentence structure	12.	Short sentences
Semantics	13.	No negation
Text	14.	No lexical variation in the text: same designation for same concept
	15.	Relevant information first
	16.	Clear structure: subheadings are used
	17.	Readers are addressed directly

Fig. 1: Rules common to all three sets of practical guidelines (BREDEL & MAAß, 2016, p. 22)
English version: (MAAß, Easy Language - Plain Language - Easy Language Plus. Balancing Comprehensibility and Acceptability, 2020, p. 75)

2 Data sources and volumes

For analysis and comparison of the application of five essential German Easy Language rules, a corpus formed by Austrian Easy Language texts and another corpus formed by Brazilian Easy Language texts were created.

For the Austrian corpus, the texts contained in it were collected from the following freely available sources:

- Österreichisches Parliament (https://www.parlament.gv.at/)
- Nachrichten leicht verständlich (www.apa.at)
- Leichter Lesen (www.sozialministerium.at)

- Bericht Unabhängiger Monitoringausschuss 2020 (https://www.monitoringausschuss.at/download/berichte/MA_Bericht_BBB_LL_2021.pdf)
- Leichter Lesen – BMJ (https://www.bmj.gv.at/service/Leichter-Lesen.html)
- Meine Rechte beim Wohnen (www.land-oberoesterreich.gv.at)

All sources for Austria were translated and/or created by Capito.

Regarding the Brazilian corpus, it was more difficult to find contents written in Easy Language. In Brazil, public institutions and private companies in different states have been working on their own projects for the creation and employment of Easy Language and the rules for it. Additionally, and different from Austria where most websites offer a specific Easy Language button which then takes the user to the corresponding content written in Easy Language, Brazilian online contents written in Easy Language are not clearly identified as such and, consequently, it was not possible to collect data as easily as it was from Austrian websites without further information and/or analysis. In spite of that, thanks to the Brazilian Network of Easy Language, Easy Language project manuals containing links and references to websites already available in Easy Language were obtained.

The corpus is set up as follows:

- Annual report of the Brazilian National Water and Sanitation Agency (ANA) from 2020 (https://www.gov.br/ana/pt-br/centrais-de-conteudos/publicacoes/carta_relatorio_ana_2020_v6.pdf)
- Coronavirus prevention manual
- (http://www.movimentodown.org.br/wp-content/uploads/2021/10/GuiaVacinacaoECuidados.pdf
- Public notice of the state of Ceará regarding citizenship and cultural diversity 2022 (http://editais.cultura.ce.gov.br/2022/02/16/edital-ceara-da-cidadania-e-diversidade-cultural/)
- Ombudsman webpage of the Audit Court of the state of Santa Catarina (https://www.tcesc.tc.br/ouvidoria#, only accessible with the use of a Brazilian VPN)

These specific contents were elaborated by different institutions in three Brazilian states (Rio de Janeiro, Ceará, Santa Catarina) and by one federal institution (National Water and Sanitation Agency). It was particularly interesting to receive contents from and to witness the engagement in the creation and employment of Easy Language by the state of Ceará. According to the Synthesis of Social Indicators (Síntese de Indicadores Sociais, SIS) by the Brazilian Institute of Geography and Statistics (Instituto Brasileiro de Geografia e Estatística, IBGE), the state situated in the Northeastern region of Brazil registered 9.3% of extremely poor people and 40.6% of poor people in 2020 (IBGE, 2020) resulting in a correspondingly poor access to education and an increased demand to meet the special needs of the affected population regarding information accessibility.

In order to find similar contents for both countries, the texts were selected accordingly. In terms of legal contents, there were collected relevant data from the Austrian Social Ministry as well as from the website of the state of Higher Austria due to the fact that the Austrian Ministry of Justice, which includes the State and District Courts, does not yet offer its website in Easy Language.

The corpus for Austria contains 14,896 tokens, and the Brazilian corpus, 14,839.

3 Description of the German rules analyzed

In order to determine and evaluate the compliance with the consolidated German rules in the Easy Language texts utilized in both corpora, five rules were selected, and their application was individually analyzed for each country. The five rules chosen are as follows:

3.1 Average sentence length

One of the basic Easy Language rules is the creation of simple sentences that convey only one idea at a time. That means that long sentences have to be divided in several sentences in order to avoid the use of subordinate clauses that not only make a sentence more difficult to read, but also modify the meaning of the main clause.

3.2 Separation of compound words with hyphens

Another rule that applies to German Easy Language is the separation of compound words with hyphens. Contrary to the traditional German orthography, words that normally do not contain a hyphen are hyphenated to make the sequence of letters not only better readable, but also to maintain the image of the different words contained and thus avoid incorrect interpretation. This can be helpful for people with reading difficulties when considering, for example, that many compound nouns in German receive an “s” between them, e.g., “Gesundheitsamt” (health department), whereas the separate nouns in this example are “Gesundheit” (health) and “Amt” (department). Therefore, a long term such as “Mühsamkeit” (strenuousness) should be separated into its identifiable word parts (“Müh-sam-keit”). It is important not to split a word into its syllables because syllables are not always meaningful to the reader as they often do not coincide with the word boundary.

Especially in Germany, there is a discussion on which punctuation to use for the separation of long or compound words: The dispute is about whether to use the hyphen (“-”) or the interpunct (“·”) specifically called “Mediopunkt” (literally: “mediopunct”) for Easy Language. The main critique of the hyphen is:

– that it may not always be semantically clear: “Markt-Führer” (market leader) may be understood as a market leader but in German could also very well be understood as a guide that leads one through a market;

– that it is more difficult to read than the interpunct after all, as the first letter after the hyphen should be written in uppercase: the German word “Herzversagen” (heart failure) would be written as “Herz-Versagen” instead of “Herz·versagen” with the interpunct;

– that individuals that are in the process of learning German through Easy Language will learn and get used to employing the hyphen where in fact there would not be one and where it is orthographically incorrect in German.

An alternative to this decision is to use both: the hyphen where it is orthographically correct in German, and the interpunct for the other cases where needed as auxiliary character for reading (MAAß, 8. Mediopunkt statt Bindestrich, 2014).

3.3 Avoidance of subordinate conjunctions

This rule relates to the average sentence length and the importance of not modifying further the idea of a main clause. According to the rules for Easy Language published by the German Network for Easy Language (NETZWERK LEICHTE SPRACHE), the conjunctions “oder” (or), “wenn” (when/if), “weil” (because), “und” (and), and “aber” (but) may be used in the beginning of a new sentence after splitting two clauses into two sentences:

“Bitte gehen Sie nach Hause oder rufen Sie ein Taxi.”

(Please go home or call a cab.)

Possible solution:

“Bitte gehen Sie nach Hause.	*(Please go home.)*
Oder rufen Sie ein Taxi.”	*Or call a cab.)*

When examining the same rule in Duden (BREDEL & MAAß, 2016), the authors advise against doing so, especially when it comes to conditional subordinate conjunctions because they are often taken out of context by transforming the main or subordinate clause into a second sentence in a new line and are thus prone to misinterpretation of the meaning.

“Patienten müssen das Krankenhaus nach einer Operation verlassen.
Wenn sie nicht mehr auf stationäre Betreuung angewiesen sind.”

(Patients have to leave the hospital after a surgery.
When they no longer need inpatient care.)

If the reader does not take into consideration that the first sentence is a condition of the second sentence, he or she will understand that all patients have to leave hospital (right) after surgery.

For causal and final subordinate clauses, Duden recommends splitting the subordinate from the main clause, inverting them and replacing “weil” (because) or “damit” (so that) with “deshalb” (therefore) in the beginning of the new line:

“Ich muss nach Hause gehen, weil ich den Papagei füttern muss.”

(I have to go home because I have to feed the parrot.)

Possible solution:

“Ich muss den Papagei füttern.
Deshalb muss ich nach Hause gehen.”

(I have to feed the parrot.
Therefore, I have to go home.)

Similarly, “obwohl” (although) in concessive clauses may be replaced with “trotzdem” (nevertheless/in spite of this). Other constructions like consecutive clauses do not present a standard solution. And for temporal clauses it may be necessary to deviate from the natural course of time and follow the text type, i.e., instead of beginning with the result of an incident such as a news text would describe a car accident, it is better to maintain the news style but at the same time rearrange the chronological order.

3.4 Avoidance of personal third-person pronouns

The reiteration of textual objects in form of personal pronouns, e.g., “der Vater” … “er” (the father … he), can cause difficulties in receiving the meaning for people with reading disabilities. The reader has to refer back to something that has been said before. In addition, it is possible that the use of pronouns is not entirely reliable and clear as seen in the example below:

“Susi geht mit Nina wandern. Sie ist sich nicht sicher, ob sie die richtigen Schuhe dafür trägt.”

(Susi goes hiking with Nina. She is not sure if she is wearing the right footwear.)

Is it Susi who is not sure about her footwear or is it Nina?

An additional problem specific to German is that the pronoun “sie” has three different meanings: she, they, and the formal “you” (“sie” written with an uppercase initial).
For all the above-stated reasons, pronouns of the third person, in particular those referring back to a previously mentioned person or object, should be avoided. On the other hand, it is recommended to directly address readers when the person that is asked to take an action is not entirely clear (e.g., “man” (one), “jemand” (someone)).

3.5 Use of bullet characters for clear structuring

When it is necessary to link several coordinate clauses or parts of them, it is recommended to use vertical lists with bullet characters. These should consist of a title followed by the indented linked phrases:

"Haben Sie sich schon überlegt, wie Sie Ihr Geld anlegen? Möchten Sie es auf ein Sparkonto einzahlen? Möchten Sie in Aktien investieren? Möchten Sie eine Immobilie kaufen?"

(Have you already thought about how to invest your money? Would you like to deposit it into a savings account? Do you want to invest in stocks? Do you want to buy a property?)

Possible solution:
"Wie lege ich mein Vermögen an?

- *Zahle ich es auf ein Sparkonto ein?*
- *Investiere ich es in Aktien?*
- *Kaufe ich eine Immobilie?"*

4 Calculations and applied methods

The corpora were created by either copying the texts directly from websites or converting texts in PDF format into an editable format with Abbyy FineReader.

To verify and compute the implementation of the rules set out above, the author created a short Python script (see Appendix). The following libraries were used:

- **re** The python standard library "re" is used to work with regular expressions. For these corpora, the library was applied to remove e-mail addresses and urls as these affected the count of long/compound words in both.
- **Counter** The data structure "Counter" is part of the "Collections" built-in Python module and is used for different container types. For this assignment, the dictionary subclass "Counter" was used to facilitate counting of the relevant elements, namely the individual subordinate conjunctions and the pronouns.
- **spacy** "spacy" is a free and open-source library in Python used for Natural Language Processing. This library is not a standard library and thus has to be downloaded and installed for the respective language(s) in the integrated development environment (IDE). For this assignment, "spacy" was used to capture subordinate conjunctions as well as personal pronouns. For German, spacy offers fine-grained POS tags to break down pronouns into only personal pronouns. This was very useful as it was easy to filter out the relevant third-person pronouns. For Portuguese, this granulation is not yet available, and the personal pronouns used in this study were thus extracted manually from pronouns in general.

The code was created in Pycharm and is abundantly commented in the Appendix. Besides the use of the above-mentioned libraries the code was written in the course of elaborating and writing this assignment as deemed suitable rather than basing the whole assignment on a script with beginning and end – as it would have been for an explicit programming task. Therefore, in its current state, the script is not suitable for larger corpora due to a great number of for-loops used for calculating the respective items which result in a longer computing time.

5 Results, interpretation and discussion

5.1 Average sentence length

The existing German and Brazilian Easy Language rules are very clear about the importance of transforming long and complex sentences into short ones that convey one idea only, as it enhances understanding and receiving the message. The computed average sentence length for German is 5.92 words and, for Brazilian Portuguese, 5.34 words. The fact that there are no compound words in Portuguese compared to the quite frequently occurrence of them in German results in several Portuguese words for one German word, e.g., "social security contributions: "Sozialversicherungsbeiträge" vs. "contribuições para a segurança social". At a first glance, this 1:5 ratio would suggest a higher word per sentence count in Portuguese; however, the German sentence length is slightly higher. This can be explained by the fact that in Portuguese personal pronouns with subject or predicate function (pronomes pessoais do caso reto) can be omitted, as the verb itself indicates number and person:

Wir haben nichts zu essen.	*Temos nada para comer.*

In the example above, the ending "-mos" in the verb "temos" already indicates the subject as plural first person.

Another explanation can be found in the corpus itself: The materials in Portuguese were mostly obtained from websites whereas the German corpus, in proportion, contains more reports which tend to have a more consistent structure than webpage content. Therefore, when examining both corpora, it can be seen that the sentence lengths vary a lot more in Portuguese than they do in German.

Regarding the number of sentences, the German corpus contains 1,985, and the Brazilian corpus, 1,646. The reason for this difference of roughly 340 lines is rather related to design than to linguistic features. When examining the Brazilian corpus, it becomes apparent that, in comparison to the selected texts for German, there are far more single words in the Portuguese materials, either as links to new webpages or links to topics or in the form of keywords, which in the German materials are more frequently included in short and explanatory sentences. A second explanation is that the Portuguese corpus contains the annual report of the Brazilian National Water and Sanitation Agency which is highly technical by nature. Although most of its parts obey the short sentence rule, this specific text contains a few exceptions with sentences of over 40 words.

5.2 Separation of compound words with hyphens

In order to understand the relation of long words in both languages, words with over 12 characters were counted. In the Austrian corpus, 1,108 were found, and in the Portuguese corpus, 423.

As for Austria, the application of the hyphen to separate compound words proved to be used throughout the corpus, but in two different ways. In most occurrences the employment of the hyphen shows to be consistent with the rule (e.g., "Dritt-Impfungen" (third vaccinations), "Pensions-Gutschrift" (pension credits)). In some cases, however, nouns with two components were not separated, such as "Sozialministerium" (Ministry for Social Affairs). When searching the corpus for this word, occurrences with a third compound were found, such as "Sozialministerium-Service" (Service of the Ministry for Social Affairs). These words were separated but only after the second compound, which does also not correspond to the German rule for separation of compounds.

In Portuguese, compounds of two or more nouns are not joined as in German, instead they are written from back to the front, and divided by the preposition "de" (similar to the English preposition "of"), which explains the very few occurrences of long words. The following example shows the translation of the compound noun "steamship travel company":

Dampfschifffahrtsgesellschaft	*Sociedade de passeios de navios a vapor*

In the example above, "Dampfschiff" (navio a vapor) is separated from "Fahrt" (passeios) and "Gesellschaft" (sociedade) in Portuguese.

Of course, this rule applies specifically to German. On the other hand, there do exist quite long words in Portuguese, too, especially those with Greek or Latin origin prefixes, such as “autoestima” (self-esteem) or “heteroidentificação” (hetero-identification). The New Orthographic Agreement (Reforma ortográfica : guia de bolso, 2009), signed in 1990 with other Member States of the Community of Portuguese Speaking Countries (Comunidade de Países de Língua Portuguesa, CPLP) to standardize spelling rules, was ratified by Brazil in 2008 and implemented (without obligation) in 2009. The deadline for adoption in Brazil was December 31, 2015, according to Decree 7875/2012. Now before the adoption of the agreement, one of the uses of the hyphen was when the prefix ends in a vowel other than the vowel at the beginning of the second element (e.g., driving school: “auto-escola”). This no longer applies. To make matters worse in terms of Easy Language, in cases where the prefix ends in a vowel and the second word begins with “r” or “s”, these letters are now doubled, and the hyphen is not used anymore: the former “auto-suficiência” (self-sufficiency) became “autossuficiência”, “contra-senso” (nonsense, paradox) became “contrassenso” and “anti-rábico” (anti-rabies) became “antirrábico”, just to name a few.

In order to avoid these hard-to-read and hard-to-fathom words, it could be a good solution to follow the German rule of separating compounds by a hyphen or an interpunct (“auto·suficiência”) for Easy Language contents only, and adapt the rule for compounds of nouns to cover compounds consisting of prefix and noun in Portuguese.

5.3 Avoidance of subordinate conjunctions

With regard to subordinate conjunctions, the focus was given to the three conjunctions with the highest occurrence for each language. With a total of 329 conjunctions in German and 259 in Portuguese, the percentage of subordinate clauses accounts for 16.57% and 15.74% in relation to all sentences, respectively. In German, “wenn” (if/when) occurs 135 times. It is exclusively used conditionally. The conjunction “dass” (that) occurs 72 times, 9 times of which it introduces a bullet point. Subordinate clauses beginning with both “wenn” and “dass” amount to 10.5% of all sentences which is quite a high percentage in relation to all sentences. When analyzing the corpus, it can be observed that apparently, in the Austrian corpus, the use of the two conjunctions is not only acceptable but also relatively frequent when creating or translating into Easy Language. However, it is important to mention that these subordinate clauses are divided by their main clauses into a new line after between 50 and 60 keystrokes and after setting a comma, which enhances understandability and reduces the aforementioned risk of misinterpretation. “Damit” (so that) ranks third, with 33 occurrences. This conjunction could be easily replaced with “deshalb” as recommended by the German rule:

“Am Corona-Virus erkrankte Menschen sollen das Haus nicht verlassen,
damit sie das Virus nicht weiter verbreiten.”

(People suffering from the corona virus should not leave their homes
so that they do not spread the virus further.”)

The above example could be transformed into:

“Am Corona-Virus erkrankte Menschen können das Virus weiterverbreiten.
Deshalb sollen sie das Haus nicht verlassen.”

(People suffering from the corona virus can spread the virus further.
Therefore, they should not leave their homes.”)

Adding to this example, “Am Corona-Virus erkrankte Menschen” is a very difficult syntactic construction of a subject with a restrictive attribute containing preposition and verb in past participle used as an adjective. A simpler and more consistent approach with Easy Language would be:

“Corona-Kranke können das Corona-Virus weiterverbreiten.
Deshalb sollen Corona-Kranke das Haus nicht verlassen.”

(Covid patients can disseminate the coronavirus.
Therefore, Covid patients should not leave their homes.)

In Portuguese, the three highest ranking subordinate conjunctions are “que” (that), “se” (if/when) and “como” (as, how) with 35, 32, and 16 occurrences, respectively. There are other words with high results found by spacy, e.g., “para” (for), “a” (to) or “de” (of). However, these words are prepositions and not conjunctions, or they miss the word “que” to be considered a conjunction, such as it would be the case of “para que” which then means “so that”. 20% of the time “que” is used as a content clause which is not as problematic as the 80% of “que” used as relative pronoun introducing subordinate clauses that function as an attribute and, additionally, present the aforementioned difficulty of back referencing to the main clause. The latter should be investigated as to find other solutions and obtain fewer long sentences. As in German, “se” is for the most part used conditionally with one exception where it is used as “whether”. Although conditional clauses come very handy, especially when thinking of texts that contain instructions that depend on the circumstances (as it is the case of a large portion of the texts contained in the Portuguese corpus), it would be recommendable to find an alternative, perhaps as indicated by the German rules. The conjunction “como” poses fewer problems, as 60% of all identified occurrences were in fact used as the modal verb “as” and only 40% as “how” or “how to”. Unlike in German where we always need a pronoun, especially “how to” is a good solution in Portuguese for explanations (similarly to English because “como” is simply followed by the infinitive), like in this corpus: “Como acompanhar o serviço?” (How to follow up on the service?) or “Como prevenir o coronavírus” (How to prevent coronavirus). Generally, the use of subordinate clauses might need to be handled differently from the German rule for two reasons: Firstly, the personal pronoun can be omitted thus reducing sentence length; and, secondly, commas are not as often used in Portuguese between main and subordinate clauses as they are in German, which enhances the visual perception of a sentence making it more compact.

5.4 Avoidance of personal third-person pronouns

The general use of personal pronouns is as follows:

German			**Portuguese**		
	Pronoun	**Count**		**Pronoun**	**Count**
She/they/her (direct) /them(direct)/you (form.)	Sie/sie	**473 (93% vs. 7%)**	*She/they (fem.)*	ela/elas	**5 (60% vs. 40%)**
He	er	**13**	*He/they (masc.)*	ele/eles	**12 (67% vs. 33%)**
It	es	**210**	*You/him/her/them (all indirect)*	lhe/lhes	**0**
Him (indirect)	ihm	**1**	*Him (direct)*	o	**39**
Her (indirect)	ihr	**2**	*Her (direct)*	a	**-**
Him (direct)	ihn	**1**	*Them (direct, masc.)*	os	**-**
you (indirect form.)/them (indirect)	Ihnen/ihnen	**24 (92% vs. 8%)**	*Them (direct, fem.)*	as	**1**

For the analyzed corpora, this assignment focuses on the use of he/she (and “it” for German) as the other pronouns do not appear or only appear in a low count, except “Ihnen/ihnen” in German and “o” in Portuguese. The Portuguese pronoun “o” and the pertaining feminine and plural forms (“a”, “as”, “os”) were analyzed because they are identical to the four existing definite articles and the most difficult-to-recognize pronouns in Portuguese. So, in addition to having to refer back to a person or object in the text, it may be difficult for readers to distinguish between the pronoun and article use of these tiny but very common four words.

The use of the pronoun “Sie” and “sie” in German accounts for 3.2% in relation to the whole token count. However, the German corpus texts are of explanatory and advisory nature which justifies the use of the formal “you” in German to make it clear who is being addressed. The pronouns “er” (he) and “sie” (she) are in fact only used 0.3% as to refer back to a person, and appear six times together (gendering). “Ihnen” and “ihnen” occur 24 times in the corpus. But this can also be justified because when written in uppercase the pronoun in its indirect form addresses the reader directly, e.g., “Ich helfe Ihnen.” (I help you). The pronoun “es” (it) is, in fact, quite rarely used to refer back to an object or an element, but mostly together with “there” as “there is”. “Es gibt” (there is) is very informal and most of the times students are taught from very early not to use it as it is not recommended stylistically. For Easy Language purposes, however, it is definitely a simple alternative for descriptions or explanations.

In Portuguese, “you”, formal and informal, does not represent the same issue as in German due to the fact that “você” (73 occurrences) cannot be confused with any other pronoun as it is unique in pronouns, except its plural form “vocês” (1 occurrence). The reason for the quite lower count of “you” in Portuguese is that a large part of the corpus is composed of a technical annual report that mostly does not refer specifically to the reader. In the instructional texts contained in the corpus, there is mostly used the imperative verb form where, again, there is no need of adding the personal pronoun in Portuguese. “Ele” (he), “eles” (them masc.), “ela” (she) and “elas” (them fem.) are only represented in 0.11% of all tokens and can therefore be ignored. Regarding the direct male third-person pronoun “o”, the author of this study verified all occurrences manually. In fact, the pronoun appears only once in “… usuários notificados a **enviar dados** que não **o** fizerem ...” (users notified to send data who do not do this, where “this” refers to “send data” and is represented in Portuguese with the direct male singular pronoun “o”). As previously mentioned, spacy does not yet offer fine graining in Portuguese. The fact that “o” is also the male article in Portuguese somewhat explains the difficulty of it being recognized by spacy as such and not counted. Furthermore, it is interesting that the same did not happen with the female counterpart “a/as”: the one and only occurrence of the plural direct pronoun “as” in “organizando em **categorias** como **as** que estão no quadro a seguir” (organizing into categories like the ones in the table below, where “the ones” refer back to the categories and are represented in Portuguese with the direct female plural pronoun “as”) was correctly identified by spacy. Therefore, the recognition issue with “o” definitely demands further investigation.

5.5 Use of bullet characters for clear structuring

The total number of bullet characters in the German corpus amounts to 227 accounting for 11,44% of all sentences in the corpus. For the Brazilian corpus, the count of bullet characters totals 260 accounting for 15,59% of all lines. Also here, the higher number in Portuguese can be explained by the difference in the contents collected from more websites in Portuguese and from more reports in German. Nevertheless, the numbers obtained confirm that in both countries bullet characters are used to structure the texts and provide a better and easy-to-understand overview, which can additionally be verified in the collected material.

6 Conclusion

This assignment focuses on five essential rules for the creation of or translation into Easy Language. These rules were elaborated in Germany and were used in this study because it is the most comprehensive set of rules in both German and Portuguese. The objective was to explain these five rules and then verify compliance with them in Easy Language contents created in Austria and in Brazil.

The average sentence length is quite similar in both languages. For Portuguese, and upon manual verification, the sentence lengths presented a higher variation due to the composition of the corpus, which contained one highly technical text. Generally, average sentence lengths for German and Portuguese of 5.92 and 5.34 words, respectively, show that in both countries the creation of short and concise sentences is considered a central rule to contribute to easy-to-read texts.

By testing the compliance with the separation of compounds rule, it was established that in the Austrian texts collected, the rule is only followed partly, i.e., the hyphen is only used consistently between the second and a third compound. For Portuguese, as compound nouns are not used as in German, it could still be considered to use a similar rule for longer words, especially the ones consisting of a Latin or Greek prefix and noun.

Subordinate conjunctions are quite frequently used in both languages. In Austria, this is solved with splitting the subordinate or the main clause, whichever comes last, into a new line in order to maintain good visibility and (apparently) short sentences. In Brazil, however, the use of conjunctions (when not introducing a new idea to the sentence) is in most cases less disadvantageous, as it is not only possible to omit pronouns in subordinate clauses which results in shorter sentences, but also due to the fact that in Portuguese commas between main and subordinate clauses do not have to be applied as mandatorily as in German. Thus, in Portuguese even a longer sentence with main and subordinate clause does not necessarily contain the in-sentence-break that a comma usually induces.

With regard to third-person pronouns, the Austrian texts showed a low usage, except for the formal you “Sie”, indirect “Ihnen” and for “es” being used in “there is”. In the Brazilian corpus, none of the correctly calculated pronouns stand out, which is mainly explained with the fact that pronouns can be omitted in Portuguese, including in the imperative tense. Therefore, not even “you” was used very often. The incorrect count of the Portuguese third-person male direct pronoun is definitely an interesting question to be answered in the future.

Lastly, the use of bullet characters was analyzed. The employment of bulleted sentences was extremely similar, suggesting that for both countries this rule, as well as the abovementioned short sentence rule, seems to make sense in the respective languages as a means to structure texts better.

For the corpus created and analyzed, the rule both corpora complied with the most, when looking at the resulting numbers, is the use of bullet characters. They are used in both languages to simplify reading and the overall view in order to enhance receiving the message. The German rule to avoid third-person pronouns was also applied consistently in both the Austrian and the Brazilian corpus. For the rule on simple sentences, the texts in the Austrian corpus were created with short sentences and a very consistent average number of words per sentence. In Brazilian Portuguese, the average sentence length was also complied with, except in the technical annual report where the average number of words per sentence was frequently significantly higher. With regard to long words, compounds were not separated consistently in the Austrian corpus as recommended by the German rule. In the Brazilian corpus, this rule was not yet applied, but it could be in the future as in Portuguese, too, there exist long words; however, it would not be used for compound nouns, but rather in nouns with Roman and Greek prefixes. The rule that was least complied with in general and in both languages was the avoidance of subordinate conjunctions. Not only for the contents of the Austrian corpus, but also for those of the Brazilian corpus, easier-to-read texts could be achieved by creating sentences of subject-verb-predicate structure only.

The increasing availability of online material opens up a broad range of possible future work. Firstly, it would be very interesting to follow up on and collect further content created in Easy Language in both countries and not only repeat, but also extend analyses on further rules. Another (more technical) future approach is to elaborate a complete Python script according to previously defined specific calculations, which is also suitable for larger corpora. Therefore, it would also be interesting to undertake further investigations on corpus cleaning addressing the peculiarities specific to Easy Language.

Acknowledgements

I would like to thank Professor Silvia Hansen-Schirra, Ph.D., for being my supervisor on this assignment and providing very helpful resources.
I also want to thank Heloisa Fischer, M.A., who made it possible for me to get access to the latest Easy Language projects all over Brazil.
Finally, I want to thank Renate Fischer for providing information about Easy Language in Austria.

References

AÇÃO EDUCATIVA; INSTITUTO PAULO MONTENEGRO. (2018). Retrieved from Indicador de Alfabetismo Funcional (Inaf): https://acaoeducativa.org.br/wp-content/uploads/2018/08/Inaf2018_Relat%C3%B3rio-Resultados-Preliminares_v08Ago2018.pdf

BARBOZA, E. (2010, jul./dec.). A linguagem clara em conteúdos de websites governamentais para promover a acessibilidade a cidadãos com baixo nível de escolaridade. *Inclusão Social*, pp. v. 4 n. 1, p. 52-66.

BIBLIOTECA DIGITAL DA CÂMARA DOS DEPUTADOS. (2009). *Reforma ortográfica : guia de bolso.* Brasília: Câmara dos Deputados, Edições Câmara. Retrieved from Reforma ortográfica - Guia de bolso: https://bd.camara.leg.br › reforma_ortografica

BREDEL, U., & MAAß, C. (2016). *Leichte Sprache - Theoretische Grundlagen, Orientierung für die Praxis.* Berlin: Dudenverlag.

FREYHOFF, G. (1998). *Make it simple : European guidelines for the production of easy-to-read information for people with learning disability for authors, editors, information providers, translators and other interested persons.* ILSMH European Association.

IBGE. (2020). *Instituto Brasileiro de Geografia e Estatística*. Retrieved from https://www.ibge.gov.br/

MAAß, C. (2014). Retrieved from 8. Mediopunkt statt Bindestrich: https://www.uni-hildesheim.de/media/fb3/uebersetzungswissenschaft/Leichte_Sprache_Seite/Publikationen/Antworten_zu_Leichter_Sprache__Forschungsstand/8._Mediopunkt.pdf

MAAß, C. (2020). *Easy Language - Plain Language - Easy Language Plus. Balancing Comprehensibility and Acceptability.* Berlin: Frank & Timme.

MENDONÇA, N. (1987). *Desburocratização linguística: como simplificar textos administrativos.* São Paulo: Pioneira.

NETZWERK LEICHTE SPRACHE. (n.d.). *Die Regeln für Leichte Sprache.* Retrieved from https://www.leichte-sprache.org/wp-content/uploads/2017/11/Regeln_Leichte_Sprache.pdf

OECD. (2016). *Skills Matter: Further Results from the Survey of Adult Skills.* Paris: OECD Publishing.

Reforma ortográfica : guia de bolso. (2009). Brasília: Câmara dos Deputados, Edições Câmara.