An Analysis of Morphosyntactic Preferences of L2 Esperantists

Daan Brugmans (s1080742)
Radboud University, Netherlands
daan.brugmans@ru.nl

ABSTRACT

This paper presents an analysis of word order preferences in texts written by L2 Esperantists. An Esperanto treebank was analyzed to calculate distributions of word order for subject, object, and verbs in Esperanto texts from authors of L1's with varying dominant word orders. Word order use in Esperanto texts of authors with an SVO, SOV, SVO & SOV and VSO L1 were analyzed. It was found that no major differences exist between L2 Esperantists' word order use when grouped by L1 dominant word order. The findings corroborate prior research on the topic and suggest additional research possibilities for the topic.

1 INTRODUCTION

As recent research and development in the field of language technology has led to new advancements, an increasing interest within the field has developed regarding low-resource languages. The research and development in question include state-of-the-art deep learning architectures and language datasets of unprecedented sizes, having led to modern large language models (LLMs) that are capable of producing and comprehending natural language to a new degree. LLMs for a variety of languages have been established, but most research and development has focused on a select group of commonly used, high-resource languages, such as English.

With modern advancements on high-resource languages established, some have shifted focus towards the application of these advancements on low-resource languages. For example, LLMs pretrained on high-resource languages may be fine-tuned on low-resource languages with successful results. These fine-tuned LLMs may contribute towards the understanding, preservation, development, and spreading of minority or endangered low-resource languages.

One such low-resource language is Esperanto, a constructed language developed by Zamenhof [10]. Zamenhof created Esperanto with the aim that it could serve as a neutral second language, not bound to a nation or group of people, that could be spoken interculturally, while being relatively easy to learn and use. While Esperanto was adopted by many Europeans swiftly after its inception in the late 19th century, this popularity has waned with time, with English increasingly being used as a lingua franca instead. This development is reflected in the fact that nowadays, Esperanto is a low-resource language within the field of language technology.

This paper builds on top of research performed by Bick [2] and analyzes morphosyntactic preferences present within Esperanto texts. It aims to analyze the ways in which an author's first language (L1) influences the way they write in a second language (L2). Specifically, this paper aims to corroborate or reject findings about word order use of subject and object in Esperanto depending on the author's native language's dominant word order. The research question of this paper is as follows: *Does the dominant word order of an L2 Esperantist's L1 influence the word orders used when writing in Esperanto?*

2 RELATED WORK

In his book *Fundamento de Esperanto* [10], Zamenhof provides the official standard of Esperanto grammar and vocabulary. However, this work is dated, as Esperanto has evolved over time. Koutny [6] provides a concise overview of modern Esperanto phonetics, morphology, syntax, semantics, and pragmatics.

Studying differences of L1 authors' and L2 authors' texts, Lu & Ai [7] researched syntactic differences in college-level argumentative essays between L1 writers of English and a variety of L2 writers of English. Their findings were statistically analyzed using 14 different measures of syntactic complexity. From the statistical analysis, they found that, when grouped together, L2 writers showed a lower degree of phrasal sophistication (measured by complexity of nominals) and mean clause length. When considering the L2 groups independently, they found that there were clear differences between syntactic complexity of essays between language groups. For example, German L1 writers showed a generally higher level of syntactic complexity when compared to English L1 writers, while Japanese and Chinese L1 writers showed a generally lower level of syntactic complexity. Although Lu & Ai could not draw the conclusion that there exist specific causal relationships between L1 features and L2 syntactic complexity, their work corroborates the presumption relevant to this paper that the specific L1 of an author of a text influences syntax used in their L2 texts, meaning that specific L1's may have specific influences on written Esperanto syntax.

Within the overlapping fields of Esperanto and Language Technology, Eckhard Bick's work is responsible for important relevant advancements in the field and has allowed for an increased range of possibilities when it comes to Esperanto Language Technology research.

Bick developed EspGram [1], a Constrained Grammar (CG) parser for Esperanto. EspGram is capable of automatically tagging Esperanto corpora with morphosyntactic information, such as part-of-speech and syntactic function. Bick evaluated EspGram on a corpus of roughly 18.5 million words from a variety of sources, such as literature, news, and Wikipedia articles. The evaluation showed that EspGram could reach an accuracy of 99.5% when tagging part-of-speech and 92% when tagging syntactic function.

The development of EspGram has allowed the automatic morphosyntactic annotation of Esperanto corpora. Using EspGram, Bick developed a dependency-graph treebank for Esperanto called Arbobanko [2]. Arbobanko consists of a corpus of circa 50.000 words sampled from the Esperanto-written news journal Monato

TxMM'23-'24, Radboud University, Nijmegen, Netherlands

[1] that was first automatically annotated using EspGram, and then manually corrected where necessary. All words in Arbobanko are annotated with a lemma, part-of-speech, inflection, syntactic function, and dependency-head id's. In his work, Bick states that, in the future, additional annotations will be supplied to Arbobanko, such as a semantic type tag for categorizing content words semantically and morpheme structure.

With the construction of the treebank, Bick analyzed the texts of Arbobanko linguistically. In his analysis, he researched if Esperanto's free word order could be corroborated or rejected in practice, and concluded that Esperanto could be classified as an SVO (Subject-Verb-Object) and ADJ-N (Adjective-Noun) language. This conclusion is based on the finding that, for all sentences in the Arbobanko with a verb and a subject or object, 85.4% of subjects appeared left of the verb and 14.6% right, while 89.8% of objects appeared right of the verb and 10.2% left. This analysis was performed without taking the native languages of the texts' authors into account. In his work, Bick states [2] that it was yet unclear to which extent word order of Esperanto text is influenced by one's native language. It is this statement that this paper aims to expand upon.

3 APPROACH

As this paper builds on top of the work performed by Bick [2], the Arbobanko is used in order to answer the research question.

The general approach this paper adheres to is as follows: The Arbobanko provides manually-checked annotations of sentences with syntactic information, including the subject and object of a sentence. The word order of every sentence must be determined by checking the relative position of a sentence's subject and object to the sentence's verb. This is done for every sentence of every text and gives a distribution of word orders used in every Esperanto text of the Arbobanko. The texts are categorized by the author's L1's dominant word order. In order to come to that knowledge, a text's author must be known, that author's native language must be known, and that language's dominant word order must be known. When these are determined, the texts can be grouped by the dominant word order of their authors' L1, and the distribution of L2 word orders by L1 dominant word order is known. That distribution should answer the research question.

For this paper's realization of this approach, a program was developed that could automatically extract relevant information from the Arbobanko. A description of this program is given here.

First, relevant information from the Arbobanko treebank is extracted. The treebank follows the CoNLL-U format for annotating sentences with morphosyntactic information with XML between sentences that annotate metadata, such as a text's title and author. These XML tags are used to determine a text's author and the boundaries of a text within the treebank. For example, <head> tags mark the beginning of a new text. Texts are collected by iteratively analyzing sentences for syntactic information and grouping the sentences together. Once a text is completely analyzed, an author is extracted from the metadata and assigned to the text. This results in a collection of 190 texts consisting of 2,886 sentences and 52,051 tokens written by 45 different authors.

Second, the author's L1 is determined. This information is not included in the Arbobanko itself, but was manually researched for this paper. Annotations of an author's L1 and an online source supporting the provided L1 were performed in a custom generated XML file. The results of these annotations may be found in table 1. These results were loaded into the program in order to assign manually annotated L1's to the authors.

Since they are often well-known members of the Esperanto community, most authors were registered on Vikipedio, an Esperanto language version of Wikipedia. Vikipedio as a source indicates that an author's L1 can be found from their Esperanto Wikipedia entry. Monato as a source indicates that an author's L1 was determined using one of the author's articles in the Monato magazine. Other resources are included in this paper's references. If an author's L1 could not be determined, then it is marked as Unresolved and is not included in the final results.

Finally, the word orders of the texts were analyzed. This began with determining the dominant word order of every author's L1. The World Atlas of Language Structures Online [4] provides a collection of languages with annotated linguistic features, including a language's dominant word order. This paper adheres to the classifications of the World Atlas of Language Structures Online. The L1 dominant word order classifications can be found in table 2.

An L1's dominant word order is stored in the custom XML file and loaded into the program. The Arbobanko's texts are then assigned to a word order class. This word order represents the dominant word order of the language as defined by the World Atlas of Language Structures Online [4]. Then, for every L1 dominant word order, all the texts assigned to that word order class have their distribution of word orders in the Arbobanko texts calculated. This is done by determining the order in which the subject, object, and first verb of every sentence of that text appear. The result is that, for all L1's with a certain dominant word order, the distribution of word orders used in their collection of Esperanto texts are calculated. This gives the distribution of word orders used in Esperanto texts by the author's L1's dominant word order. The results are given in table 3.

4 RESULTS

Table 3 shows the distribution of L2 Esperanto word order use by L1 dominant word order. The top row shows each L1 dominant word order category. Then, for all texts of the L1 word order class, the following is shown:

- Percentage of sentences without a verb
- Percentage of sentences with a verb, but without a subject/object
- Percentage of sentences where the subject/object is left of the verb
- Percentage of sentences where the subject/object is right of the verb

These findings show that, regardless of the L1 dominant word order class, the subject appears left of the verb in a majority of sentences, while the object appears right of the verb in most sentences. This implies that the most common word order used in the analyzed texts is Subject-Verb-Object (SVO). This supports the findings of Bick [2].

[1] www.monato.be

Author	L1	Source
Josef Mendl	Czech	Vikipedio
Stefan Maul	German	Vikipedio
Paul Peeraerts	Dutch	Vikipedio
Laimius Stražnickas	Lithuanian	Vikipedio
Grigorij Arosev	Russian	Vikipedio
Bardhyl Selimi	Albanian	Vikipedio
Werner Fuß	German	Vikipedio
István Ertl	Hungarian	Vikipedio
Yamasaki Seikô	Japanese	Vikipedio
Jaroslav Klement	Czech	Vikipedio
Serge Zandandu Ntomono Zola	French	[8]
Garbhan Macaoidh	Scottish Gaelic	Vikipedio
Brian Moon	English	Vikipedio
Manfred Westermayer	German	Vikipedio
Steĉjo Norvell	English	[9]
Jean-Yves Santerre	French	[3]
Thomas Sülzle	German	Monato
Brigitte Faverial	French	Monato
Douglas Draper	Norwegian	Vikipedio
Alexander Gofen	English	Monato
Boris Kolker	Russian	Vikipedio
Elson B. Snow	English	Monato
Mu Binghua	Mandarin	Monato
Dmitrij Cibulevskij	Ukrainian	Vikipedio
Gilbert Ledon	French	Vikipedio
Thierry Salomon	French	Vikipedio
Walter Klag	German	Vikipedio
Carlo Minnaja	Italian	Vikipedio
Nikolai Gudskov	Russian	Vikipedio
Jefim Zajdman	Russian	Vikipedio
W.H. Simcock	English	Vikipedio
Hélène Falk	Dutch	[5]
Gerrit Berveling	Dutch	Vikipedio
Evgeni Georgiev	Bulgarian	Vikipedio
Ranieri Clerici	Italian	Vikipedio
Marko Naoki Lins	German	Vikipedio
Paul Gubbins	English	Vikipedio
Bertil Englund	German	Monato
Jiří Patera	Czech	Vikipedio
Hektor Alos I Font	Spanish	Vikipedio
Michael Gerenrot	Unresolved	–
Jovan Nešić	Serbo-Croatian	Vikipedio
Gonçalo Neves	Portuguese	Vikipedio
Berta Lenard	Unresolved	–
Kei Kurisu	Japanese	Vikipedio

Table 1: Authors of the Arbobanko's texts and their L1 with a source in order of appearance in the Arbobanko.

Although the subject of a sentence appears left of the verb for a majority of sentences regardless of L1 dominant word order, authors of L1 languages with an SVO word order produced sentences with a subject left of the verb 10 percentage points less than authors of other L1 word order classes. This means that, despite the fact that authors with an SVO L1 would write a subject prior to the verb naturally in their native language, they adhere to such a structure less frequently than authors with other L1 word order classes. This pattern is not seen in authors whose L1 has both SVO and SOV as dominant word orders. SVO L1 authors also tended to write the object to the left of the verb the most frequently, which also goes against the dominant word order of their native language, although here the difference between SVO L1 authors and other L1 category authors is small with a few percentage points. Aside from these patterns, the distributions of L2 word orders between L1 word order classes are noticeably similar: authors with an VSO L1, SOV L1 or SVO & SOV L1 all placed the subject to the left of the verb for roughly 75% of sentences, while placing the object to the right of the verb for roughly 50-55% of sentences.

5 DISCUSSION

The findings of table 3 do not show major word order use differences between authors of L1's with different dominant word orders. Some of the results actually seem to go against what is expected. For example, it goes against personal expectations that the group of SVO L1 Esperantists, writers who are naturally inclined to place a subject prior to a verb in their native language, place the subject of a sentence prior of the verb the least in Esperanto. However, that pattern is not found in authors with an SOV L1 or SVO & SOV L1, so this may be a bias in the data.

In addition to this bias, there are some other limitations to the findings of this paper that must be taken into consideration. A major limitation is the lack of reliable sources of authors' L1's. Although most authors of Arbobanko texts have online records, many of them do not explicitly mention their native language(s), with some authors not being traceable at all. Because of this, for varying authors included in the results of this study, there is no single concrete source to refer to, and a conclusion of an author's native language must be drawn from varying sources. Mistakes or wrong assumptions may have been made as a result.

Another limitation of the research is that there is an imbalance of dominant word orders analyzed: texts of SVO L1 authors make up over 50% of sentences analyzed, while SOV L1 and VSO L1 authors' texts have much smaller shares. This imbalance may introduce bias for underrepresented L1 word order categories, as the pool of authors for the L1 word order category may be limited and not represent a more general population of authors. A noteworthy example of this is the pool of authors for the VSO L1 category, as it only consists of a single author, Garbhan Macaoidh. Since the pool of VSO L1 texts only consists of texts by this one author, the author's personal writing style may bias the data towards a certain direction, and the resulting analysis could be interpreted not as the morphosyntactic preferences of VSO L1 writers, but of Garbhan Macaoidh.

6 CONCLUSION & OUTLOOK

It may be concluded from the findings of this paper that there does not exist a preference for L2 Esperanto word order use that is influenced by the dominant word order of an author's L1. This answers this paper's research question.

Dominant Word Order	Languages	Sentences	Tokens
SVO	Albanian, Bulgarian, Czech, English, French, Italian, Lithuanian, Norwegian, Portuguese, Russian, Serbo-Croatian, Spanish, Ukrainian	1,605	28,333
SOV	Japanese, Mandarin	170	4,261
SVO & SOV	Dutch, German, Hungarian	630	12,505
VSO	Scottish Gaelic	148	3,128
Unresolved	—	333	3,824

Table 2: Classification of author first languages by dominant word order as defined by Dryer [4]. The amount of sentences and tokens analyzed per classification is given.

L1 Dominant Word Order	***SVO***		***SOV***		***SVO & SOV***		***VSO***	
L2 Word Order	*Subject*	*Object*	*Subject*	*Object*	*Subject*	*Object*	*Subject*	*Object*
No Verb	7.48%		5.29%		3.02		4.05	
Verb without . . .	5.36%	38.75%	1.18%	31.18%	3.02%	36.51%	4.05%	41.22%
. . . left of Verb	67.04%	4.67%	77.65%	6.47%	75.87%	5.08%	74.32%	4.05%
. . . right of Verb	20.12%	49.10%	15.88%	57.06%	18.1%	55.4%	17.57%	50.68%

Table 3: Relative counts of the position of subjects and objects to a sentence’s verb, ordered by the dominant word order of the L1 of the author of the text.

Further research should take the limitations of this paper into account. I suggest that further research attempts to establish the L1 of Arbobanko authors with greater certainty by seeking, or creating, sources that are more reliably and uniform. Furthermore, a greater selection of texts from a greater selection of authors should be analyzed, so that individual authors’ writing styles cannot bias the findings. I suggest that this may be done using EspGram [1], since EspGram is capable of automatically annotating syntactic information of Esperanto text. The use of EspGram could expand the size of the data analyzed and may improve results. Finally, although this paper only focused on word order use, additional morphosyntactic phenomena may be analyzed, such as the order of adjectives and nouns. Such research could give greater insight into morphosyntactic preferences of L2 Esperantists.

REFERENCES

[1] Eckhard Bick. 2007. Tagging and Parsing an Artificial Language: An Annotated Web-Corpus of Esperanto,. In *Proceedings of the Corpus Linguistics Conference*, Matthew Davies, Paul Rayson, Susan Hunston, and Pernilla Danielsson (Eds.). University of Birmingham, Birmingham, United Kingdom. https://ucrel.lancs.ac.uk/publications/CL2007/

[2] Eckhard Bick. 2020. Syntax and Semantics in a Treebank for Esperanto. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 5120–5127. https://aclanthology.org/2020.lrec-1.630

[3] Fédération Espéranto Bretagn. 2007. Herbignac. https://bretonio.esperanto-france.org/spip.php?article79&lang=eo.

[4] Matthew S. Dryer. 2013. Order of Subject, Object and Verb (v2020.3). In *The World Atlas of Language Structures Online*, Matthew S. Dryer and Martin Haspelmath (Eds.). Zenodo. https://doi.org/10.5281/zenodo.7385533

[5] Vlaamse Esperantobond. 2020. Contactadressen. https://www.esperanto.be/fel/nl/contact.php.

[6] Ilona Koutny. 2015. A typological description of Esperanto as a natural language. *Język. Komunikacja. Informacja* 2015, 10 (january 2015), 43–62. http://jki.amu.edu.pl/files/JKI%20-%20tom%2010%20-%202015.pdf

[7] Xiaofei Lu and Haiyang Ai. 2015. Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing* 29 (september 2015), 16–27. https://doi.org/10.1016/j.jslw.2015.06.003

[8] Tresor Mavungu. 2014. Mise en place d’un systeme d’information pour la gestion des expatriés du bas Congo. https://www.memoireonline.com/10/14/8927/Mise-en-place-dun-systeme-dinformation-pour-la-gestion-des-expatries-du-bas-Con html.

[9] Memorial University. 2015. Obituary of Stevens Thompson Norvell (Jr.). https://www.engr.mun.ca/~theo/STNorvell-Obituary/ObituarySTNorvell-EN.html.

[10] Ludwik Lejzer Zamenhof. 1905. *Fundamento de Esperanto*. Hachette, Paris.

A WORK REPORT

Altough I am a Computing Scientist, I chose to lean towards a more linguistically focused research for this project, as I wanted to develop my linguistic skills. This required that I familiarized myself with some linguistic research, standards, and terminology that I was unfamiliar with. For example, I learned about the CoNNL-U format for constructing treebanks, since the Arbobanko was delivered in this format. Learning to work with treebanks was also new to me, as was exploring Esperanto research, since a lot of research about Esperanto is written in Esperanto, which I do not understand very well. For this report, I first constructed the literature section, then I took my time to understand the contents of the Arbobanko in a trial-and-error way by trying to correctly parse it in a Python program.

Only after completing the program and getting my findings, did I continue with my report. Unfortunately, I worked a lot on this project in a small period of time. If possible, for a next project, I would want to try and spread my work out more thinly across a broader time period.