The Linguistic Profile of Totalitarianism

Daniel Evans

Introduction

The goal of this project is to take an interdisciplinary look at the idea presented by Viktor Klemperer in his *LTI - Lingua Tertii Imperii*. Klemperer was a German-Jewish philologist of relatively high social class who lived in Germany under Nazi rule. In his book Klemperer provides a qualitative linguistic analysis the linguistics of Nazi rhetoric. In this paper I provide a quantitative complement to Klemperer's arguments by combining ideas from language acquisition and distributional semantics using recent developments in natural language processing. Instead of analyzing a Nazi linguistic profile, however, I analyze a Soviet linguistic profile and compare it against that of average Russian language use to extend Klemperer's arguments to totalitarian governments in general. There are some issues with making such a leap, but this project nevertheless provides some insight.

Background

People with a background in cognitive science are aware of the phenomenon where hearing, seeing, or reading something hundreds or thousands of times can make someone internalize it, consciously or not. In negative contexts, this can be referred to as propaganda, indoctrination, brainwashing or gaslighting. In the field of language acquisition, however, it's related to the main ideas of language acquisition called "statistical learning" (Saffran 2003). Statistical learning refers to how we acquire the meaning of words and phrases from repeated exposure in context. Where and when

was a word said? By who, to who? What words appeared before it and after it? We achieve a baseline understanding of language in large part by statistical learning, especially during early childhood (Saffran 1996). Over time we develop an intuition, for example, to recognize when a word is used out of place, when someone tells a lie, or when someone is making a subtle implication.

Klemperer, as a philologist, was certainly aware of this phenomenon, and actively avoided exposure to linguistic content produced by the Nazi party - posters, books, songs, speeches played on loudspeakers - to avoid being “hijacked” by Nazi ideology. The Nazi party’s ubiquitous use of propaganda through a huge number of mediums provided the average German with an abundance of rich linguistic sources. Since the language used by the Nazi party was itself different from the average individual’s language before Nazi rule, individuals had the same natural, psychological pressure to “learn” the new language that individuals in early childhood do when first acquiring language.

On the one hand, Nazi rhetoric tended toward germanolinguopurism, either favoring or creating linguistic choices of demonstrably Germanic origin. On the other, it used several words of nonGermanic origin. The Nazi party had obvious reason to favor linguistic forms that bolstered their pro-German racist narrative, but this generalizes poorly to Soviet party, which branded itself as accepting of all races. Ultimately, a totalitarian leadership’s desired effect of propaganda is not just to persuade the masses to internalize an opinion of race or the value of work, but truly to get the masses to

acquire its own intuition for its ideology by leveraging the psychological effect that endless linguistic repetition has on the human mind.

The methods used in this paper are based on an idea similar to statistical learning called “distributional semantics” - the idea that words with similar distributions have similar meanings. In short, they take a corpus of text data as input and produce statistical models that measure the semantic similarity between words. This project uses two separate models trained on different corpora of Russian language. The first is a pretrained model made available by Facebook trained on the entire Russian wikipedia and a publicly available corpus obtained from crawling the internet (Grave 2018). The second is a model trained on a collection of Joseph Stalin’s essays, letters, speeches, and other sources (Stalin 1901-1953). What makes these models particularly useful is the ability to take an input word like *bditel’nyy*‘ vigilant’ - a word found often on Soviet propaganda posters - and return a list of words that it has determined have a similar meaning. Klemperer sets the precedent for this type of signal as evidence when he describes how the Nazi party used the word ‘fanatical’, which has negative connotations, in contexts where ‘heroic’ or ‘virtuous’ would be appropriate choices in average German language.

Tools and Methods

Below are the tools used in this project:

- Python programming language: used to write the scripts to collect, process, and the Stalin corpus

- The Natural Language Toolkit, a Python library for natural language processing (Bird 2009): used for high-level preprocessing of the Stalin corpus, like determining word- and sentence-boundaries
- Gensim, a Python library for topic modelling (Řehůřek 2010): used to calculate word embeddings, mathematical objects that are used to calculate the semantic similarity between words

The collection is available online at the URL provided in the bibliography, and required an automated collection process and a large amount of preprocessing before training. Below is the procedure for collecting the corpus:

1. Find the link for each source. The collection is hosted on a website and consists of 18 volumes. Each volume has a “contents” page that contains links to separate pages for each source.
2. Remove duplicate links, since some sources do live on the same page.
3. Visit each unique link and read the page content for each source.
4. Save the content of each page to a file before any preprocessing. At this point, the contain still contains some garbage data that needs to be removed before analysis.

Below is the procedure for preprocessing the corpus and training a semantic model:

1. Remove sources with blank content.
2. Filter out the pre- and post-body content. Each item starts with some garbage data before the actual text, and ends with more garbage data.

3. Filter out page numbers. Each text may or may not have page numbers included from the original transcription.
4. Determine sentence boundaries using NLTK's implementation of the Punkt sentence tokenizer.
5. Determine word boundaries using NLTK's implementation of the Russian Toktok word tokenizer.
6. Obtain word vectors by using Gensim's implementation of FastText word vectorizer on the preprocessed text.

Since Facebook's FastText algorithm analyzes parts of words, it can learn the semantics of derivational and inflectional affixes. This brings the advantage of producing better vectors for smaller corpora and reduces the need to filter out affixes (Bojanowski 2016). The Stalin corpus contains just over 1300 sources, 88,000 sentences, and 1,500,000 words - not huge, but hopefully large enough to get some decent quality vectors.

Data

After training a model on the Stalin corpus, I make the same query to both models and compare the lists of "most similar" words they return. Here, I show comparisons made on three words commonly found on Russian propaganda posters: *bditel'nyy* 'vigilant', *narod* 'the people', and *rabota* 'work'. For clarity, I've removed words that share stems with the input words (*bditel'naya* is simply an inflected form of *bditel'nyy*, for example).

Figure 1: *bditel'nyy* 'vigilant'

Stalin	Average
reshitel'nyy 'resolute'	*nedremlyushchiy* 'awake'
dlitel'nyy 'long'	*neusypnyy* 'vigilant'
zagotovitel'nyy 'procuring'	*vnimatel'nyy* 'attentive'
podgotovitel'nyy 'preparatory'	*zorkiy* 'vigilant'
polozhitel'nyy 'positive'	*dotoshnyy* 'meticulous'
deystvitel'nyy 'valid'	*osmotritel'nyy* 'prudent'
ispolnitel'nyy 'executive'	*pronitsatel'nyy* 'perceptive'
sokrushitel'nyy 'crushing'	*nevnimatel'nyy* 'inattentive'
ob"yedinitel'nyy 'unifying'	*dobroporyadochnyy* 'respectable'
smertel'nyy 'fatal'	*dobrosovestnyy* 'conscientious'

Figure 2: *narod* 'the people'

Stalin	Average
naryad 'outfit'	*lyud* 'the people'
narkomprod 'commissariat'	*elektorat* 'electorate'
sibiri 'siberia'	*proletariat* 'proletariat'
razbrod 'disorder'	*bogonosets* 'god-bearer'
nashelsya 'be located'	*sbrod* 'rabble'
gorod 'city'	*bezmolstvuyet* 'silent'
vklad 'contribution'	*etnos* 'ethnos'
grubyy 'rude'	*okhlos* 'ochlos'

namechayet ‘outlines’	*naselyayushchiy* ‘inhabiting’
tshchatel'noy ‘thorough’	*obolvanennyy* ‘fooled’

Figure 3: *rabota* ‘work’

Stalin	Average
obrabotki ‘processing’	*kropotlivaya* ‘painstaking’
bezrabotitsu ‘unemployment’	*ucheba* ‘studies’
sabotazh ‘sabotage’	*kazhdodnevnaya* ‘everyday’
rasy ‘races’	*rutinnaya* ‘routine’
pozabotit'sya ‘take care of’	*sovmestnaya* ‘joint’
kommunizme ‘communism’	*deyatel'nost* ‘activities’
raby ‘slaves’	*tvorcheskaya* ‘creative’
slepota ‘blindness’	*prodelannaya* ‘done’
vopreki ‘contrary to’	*professiya* ‘profession’
nauchnym ‘scientific’	*vypolnyayemaya* ‘carried out’

Results and Discussion

The results are somewhat disappointing in quality, as there seem to be a lot of words considered similar due to similar spellings - for example, *bditel’nyy-dlitel’nyy* ‘vigilant’-‘long’ or *rabota-sabotazh* ‘work’-‘sabotage’ - so interpretation should be made tentatively and with care. This is very likely due to the relatively small size of the data set and the nature of the FastText algorithm. It may be the case that polysynthetic

languages such as Russian with a large morpheme-to-word ratio need a larger data set to achieve good results. However, it's still worth entertaining a few interesting results from the model outputs:

- Similar to *vigilant*: preparatory, procuring, fatal
- Similar to *the people*: commissariat, disorder, city, contribution
- Similar to *work*: communism, slaves, scientific

This exercise is an interesting quantitative approach to Klemperer's qualitative analysis of Nazi rhetoric, but due to the aforementioned issues falls short of providing convincing evidence. The next step would be to introduce higher-level linguistic preprocessing to improve the quality of the input dataset.

Conclusion

Most understand that propaganda and rhetoric are unsettling ways of spreading ideologies - they don't require a given person to agree with it initially in order to hijack people's minds regardless. It's a remarkable feat of psychological abuse that only people in power are capable of. With our understanding of linguistics, we may be able to come closer to revealing the nature of totalitarian abuse of power. Humans acquire word-meaning relationships subconsciously, and the utterly pervasive propaganda employed by totalitarian leadership essentially forces people into learning whatever word-meaning relationships the regime wants - simply by the nature of human psychology.

Bibliography

Victor Klemperer and Martin Brady (2002). *The Language of the Third Reich = LTI - Lingua Tertii Imperii: a Philologist's Notebook*. Athlone Press.

Steven Bird, Edward Loper and Ewan Klein (2009), *Natural Language Processing with Python*. O'Reilly Media Inc. (see https://www.nltk.org/)

Radim Řehůřek and Petr Sojka (2010), *Software Framework for Topic Modelling with Large Corpora*. (see https://radimrehurek.com/gensim/intro.html)

Joseph Stalin et al (1901-1953), collection of over 1300 essays, letters, speeches, and various other works. Volumes 1-13 published by the Marx – Engels Institute under the Orgburo in 1946-1952. Volumes 14-18 published by Dr. R. Kosolapov in 1997-2006. Works hosted online by the Mihail Grachev Library (see http://grachev62.narod.ru/stalin/index.htm)

Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov (2016), *Enriching Word Vectors with Subword Information*.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, Tomas Mikolov (2018). *Learning Word Vectors for 157 Languages*.

Jenny Saffran (2003). *Statistical Learning: Mechanisms and Constraints*.

Jenny Saffran, Richard Aslin, Elissa Newport (1996). *Statistical Learning in 8-Month-Old Infants*.