Universität Heidelberg

Neuphilologische Fakultät

Institut für Computerlinguistik

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

Bias-Analyse der Werke der Brüder Strugatzki

IMPLEMENTIERUNGSPROJEKT

Bias – WS2019/2020

Dozentin: Katja Markert

vorgelegt von: Denis Logvinenko

Ilse-Krall-Straße 49

69124 Heidelberg

logvinenko@cl.uni-heidelberg.de

Studiengang: Computerlinguistik (HF),

Romanistik (NF)

5. Fachsemester

Matr.-Nr: 3440438

Inhaltsverzeichnis

1 Motivation und Einleitung

1.1 Motivation

Die Werke der Brüder Strugatzki zählen zu den wichtigsten Vertretern der russischen Science- und Social-Fiction. Mit ihrem Werk haben Boris und Arkadij Strugatzki den Grundstein für die Verbreitung und Beliebtheit dieses Genre in der russischen Literatur gelegt.

Ihre Bücher widmen sich vielen gesellschaftlichen Themen, mit denen die Menschheit zu unterschiedlichen Zeiten, meistens aber in der Zukunft, konfrontiert wird. Es wird ein Versuch gewagt, mit Hilfe des Lesers folgenden Fragen auf den Grund zu gehen:

Wie wird unsere Welt aussehen? Was wird die Geister der Zukunftsschöpfer, die in der von Strugatzki erschaffenen Welt meistens einfache Menschen sind, beschäftigen *(Mittag: 22. Jahrhundert)*? Wird es einen Zeitpunkt geben, an dem sich die Menschheit in zwei Teile aufgrund der Überlegenheit des einen gegenüber dem anderen aufspaltet, und ob dann der Fortschritt allen gleichermaßen zur Verfügung stünde *(Ein Käfer im Ameisenhaufen; Die Wellen ersticken den Wind)*? Hat die aufgeklärte Erde das moralische Recht, den Fortschritt auf anderen Planeten zu beschleunigen *(Fluchtversuch; Es ist schwer, ein Gott zu sein; Die bewohnte Insel)*? Wie würde sich der Mensch verhalten, wenn er vor der Wahl stünde, entweder sein Lebenswerk aufzugeben und weiter von äußeren Umständen ungestört zu leben oder möglicherweise deswegen zu sterben *(Milliarden Jahre vor dem Weltuntergang)*? Wie würde sich eine Gesellschaft verhalten, welche man, selbst mit den besten Absichten, einem Experiment unterzieht *(Das Experiment)*?

In den Büchern selbst findet man keine Antworten – eine erschöpfende Antwort zu geben, war auch nie das Ziel der Autoren, es war ihnen vielmehr daran gelegen, Menschen in der Ära der sowjetischen Stagnation zu Reflexion zu bewegen, die während der Herrschaft von Leonid Breschnew begann. „Denkt!“ war dabei ihr Leitspruch.

Die Autoren beschreiben in ihren ersten Erzählungen und Powesti eine nahezu utopische Welt, in der es fast nichts gibt, was zu bemängeln wäre, in der alle Menschen ihren Platz oder sogar ihre Berufung im Leben umstandslos finden können, woher oder aus welcher Zeit man auch immer stammt. Erst später machen sich in ihren Werken erste Krisenvorboten bemerkbar, die ihrerseits zu vielzähligen wie oben bereits erwähnten Fragen philosophischen Charakters führen.

Wegen ihrer sprachlichen sowie inhaltlichen Komplexität heben sich Strugatzkis Werke von manchen anderen Vertretern dieses Genre positiv ab, die sich auf der Jagd nach einer immer größer werdenden Leserschaft des Öfteren der Stilmittel der Popliteratur bedienen.

Aus diesem Grund im Speziellen, sowie aus meiner Vorliebe für Science-Fiction im Allgemeinen mache ich die Analyse der Werke von Boris und Arkadij Strugatzki zum Gegenstand dieser Arbeit.

1.2 Motivation zur Analyse

Trotz der unbestreitbaren Wichtigkeit für die russische Literatur und der großen Ehrfurcht seitens des lesenden Publikums, gewinnt man bei der Lektüre dieser Werke des Öfteren den Eindruck, dass die führende Rolle sowie progressives Gedankengut in der Gesellschaft hauptsächlich den Männern gehören. Das spiegelt sich bspw. in einer geringen Anzahl an

weiblichen Charakteren wider, oder in der Art ihrer Beschäftigung, wenn sie doch im Text in Erscheinung treten. Wir verfolgen die Abenteuer der Männer, die als Angehörige der Landungstruppen, „Pfadfinder“ (die versuchen, den ersten Kontakt mit anderen Spezies herzustellen), Sternpiloten, Kapitäne, Astrophysiker, Xenopsychologen, Historiker, Leiter verschiedener Organisationen ihren Beitrag zu dem Gemeinwohl ihrer Gesellschaft aktiv beisteuern. Allerdings lässt sich die Zahl der Frauen, die vergleichbar hohe Positionen und Stellen innehaben, nur als bescheiden einschätzen. Es passiert nicht selten, dass man Frauen sieht, die nur auf die Rückkehr ihrer Männer aus langen wichtigen Expeditionen warten, Frauen, die als Dienerinnen arbeiten oder auch Frauen, deren Verhalten man milde formuliert als kurios bezeichnen kann.

1.3 Ziel und Methode

Die Fragen, denen ich in dieser Arbeit nachgehen möchte, sind folgende:

- Kann man einen möglichen Bias in den Embeddings, die auf dem aus diesen Werken erstellten Korpus basieren, quantifizieren?
- Reicht dafür dessen Größe aus?
- Lässt sich anhand dessen die Hypothese bestätigen, dass die Rolle der Frauen in der „Welt des Mittags“[1] vernachlässigt wird? Dass die in den Texten verwendete Sprache die Vorurteile und stereotypes Denken der Frauen gegenüber gut erfasst.
- Und wie lässt sich die Sprache der Werke mit den aktuell besten Modellen für die russische Sprache vergleichen?

Um die obengenannten Fragen beantworten zu können, sind folgende Schritte notwendig:

- Erstellung eines Korpus aus den Werken:
    a. Herunterladen der Texte
    b. Vorverarbeitungsschritte, die zu einem Korpus führen
- Training der Embeddings:
    a. Auswahl des besten Modells (FastText, GloVe)
    b. Tests mit verschiedenen Hyperparametern
- Bias-Analyse:
    a. Assoziation mit verschiedenen Adjektiven, die Persönlichkeitsmerkmale beschreiben
    b. Vergleich der Ergebnisse mit den aktuellen Modellen für die russische Sprache

[1] So benannten die Autoren die von ihnen erschaffene Welt.

2 Erstellung des Korpus

2.1 Zusammenstellung der Werke

Alle veröffentlichten Bücher wurden von den Verwandten der Autoren auf ihrer offiziellen Webseite [1] zur Nutzung der Öffentlichkeit freigegeben.

Die Werke lassen sich als HTM-Dateien herunterladen. Das Problem, das es in diesem Schritt zu lösen gilt, ist, dass einem nur die Hälfte aller Bücher in einem File zur Verfügung steht. Für alle anderen sind lediglich einzelne Kapitel abrufbar. Das ist jedoch nicht allzu dramatisch, da alle URLs nach einem Muster aufgebaut sind. Nehmen wir das Buch „Das Experiment“ als Beispiel:

- http://www.rusf.ru/abs/books/go01.htm – das erste Kapitel
- …
- http://www.rusf.ru/abs/books/go19.htm – das letzte Kapitel

Deshalb ist es möglich, für alle Werke die URLs von der Kapitelanzahl und der Webadresse des ersten Kapitels abzuleiten. Die Funktionen dafür (download, get_chapters) finden sich in den Teilen 0.1 bis 1.1 des dieser Arbeit beigefügten Jupyter-Notebooks. Die Tupel in Form:

(url_before_chapter_number, last_chapter_number, '.htm')

sind im Teil 0.1 definiert.

2.2 Preprocessing-Schritte

Die heruntergeladenen HTM-Files sind in dieser Form zum Training der Wort-Embeddings noch nicht geeignet und müssen entsprechend verarbeitet werden.

Die Vorverarbeitung besteht aus folgenden Schritten (alle Funktionen, die diese Aufgabe erfüllen sind im Teil 1.1 definiert):

1. Dem HTM-File muss nicht programmierbarer Text entnommen werden. Das kann mit der Bibliothek „html2text“ erreicht werden.
2. Einige Zeichen müssen ersetzt werden. Dazu kann man sich der Funktionen str.maketrans und str.translate bedienen, wobei für die erste eine Übersetzungstabelle definiert werden muss. Deshalb muss man zuerst herauszufinden, welche Zeichen (und welchen Code sie haben) das Ergebnis von html2text enthält. Für diesen Schritt wird für alle in Text konvertierten HTM-Dateien ein Set gebildet und sie werden anschließend vereinigt.
   Die Tabelle sieht folgendermaßen aus:

Zeichen	Ersetzt mit
1 2 3 4 5 6 7 8 9	0
! ? ;	.
" # % \ ' ( ) * + , / : ; = > @ [ ] _ ` | chr(147) chr(148) § « º » №	

3. Die 172 Kapitel, die mit Python heruntergeladen wurden, enthalten im Gegensatz zu den Büchern, die man ganz herunterladen kann, nicht nur literarischen Text, der für die Analyse von Interesse ist, sondern auch den Text, der auf der Webseite

Abbildung 1. Beispiel einer HTM-Datei

sichtbar ist (von Webseitenelementen wie Menüs, href-Textauszeichnungen, HTML-Tabellen usw.). Die Textdateien müssen vom Text dieser Art „gereinigt“ werden. Die Positionierung der Elemente bleibt gleich, deswegen genügt es, die ersten und die letzten N Zeilen sowie alle Zeilen, die 'htm', 'jpg', 'http', 'design' oder 'ru' enthalten, wegzulassen. Des Weiteren wird der Text anschließend in Kleinbuchstaben konvertiert.

4. Die Redakteure entschieden, manche Wörter hervorzuheben, indem das Wort mit Leerzeichen zwischen den Buchstaben geschrieben wird. Um auch solche Wörter in dem Modell nicht zu verlieren, muss man sie finden (Funktion *show_spaced_words*) und die Leerzeichen dazwischen löschen (per Hand, da es auch Fälle gibt, in welchen man ohne Kontext nicht sagen kann, ob die Buchstabenfolge ein oder mehrere Wörter enthält). Die Aufgabe wird vor allem durch den Umstand erschwert, dass es im Russischen relativ viele Wörter (Präpositionen und Personalpronomen) gibt, die aus nur einem Buchstaben bestehen:

Wort	Übersetzung
а, в, и, к, о, с, у, я	aber, in, und, zu, über, mit, bei, ich

   Alle anderen Einzelbuchstabenwörter, die zum Beispiel häufig in Namenskürzeln (z.B. М. Каммерер – M. Kammerer) vorkommen wurden gelöscht (Funktion *delete_one_letter_words*).
5. Da bei den meisten Bibliotheken/Applikationen der Input als Textfile mit einem Satz pro Zeile eingegeben werden muss, wurde der Text am Punkt '.' gesplittet (alle Satzendezeichen wurden durch '.' ersetzt).

2.3 Lemmatisierung und POS-Tagging

Die Modelle wurden auf drei Versionen des Korpus trainiert: auf dem originalen, auf dem lemmatisierten und auf dem lemmatisierten Text mit POS-Tags. Vor dem Hintergrund, dass Russisch eine stark flektierende Sprache ist (sechs Fälle, Aspekte, reiche

Morphologie) und dass man später assoziierte Wörter finden muss, wäre es sinnvoll für alle Wörter ein Lemma zu finden, mit welchem man die Ähnlichkeit berechnen kann.[1]

Für diesen Zweck gibt es die Konsolenanwendung MyStem von Yandex [2], die dem Nutzer erlaubt, grammatische Informationen in XML-, JSON-Format oder im Klartext zu speichern. Da es einfacher ist, strukturierte Informationen aus strukturierten Datenformaten zu extrahieren, wurde in dieser Arbeit für das JSON-Ausgabeformat entschieden.[2] Beispiel einer Ausgabe:

Satz (Eingabe):	нужно было только беречь пальцы[3]
MyStem (Ausgabe):	[{"analysis":[{"lex":"нужно","gr":"ADV,praed="},{"lex":"нужный","gr":"A=sg,brev,n"}],"text":"нужно"},{"analysis":[{"lex":"быть","gr":"V,intr=praet,sg,indic,n,ipf"},{"lex":"было","gr":"PART="}],"text":"было"},{"analysis":[{"lex":"только","gr":"PART="},{"lex":"только","gr":"CONJ="},{"lex":"только","gr":"ADV="}],"text":"только"},{"analysis":[{"lex":"беречь","gr":"V,ipf,tran=inf"}],"text":"беречь"},{"analysis":[{"lex":"палец","gr":"S,m,inan=(acc,pl|nom,pl)"}],"text":"пальцы"}]
Befehl:	mystem -gi --eng-gr --format json input output

Unter "analysis" findet man alle lexikalischen und grammatischen Informationen, sowie auch die Grundform des Wortes.

In der dritten Version des Korpus wurden die POS-Tags an die Wörter folgendermaßen angehängt:

Lemma_POS (z.B. "нужный_ADV")

[1] Ansonsten müsste man für jedes Wort, das man nachschlägt, auch all seine Formen auflisten und z.B. einen Durchschnittsvektor bilden.
[2] In diesem Format hat man eine list of dictionaries pro Zeile.
[3] = Man musste nur seine Finger schonen.

3 Erstellung der Wortvektoren

Ausgehend von dem im zweiten Teil erstellten Korpus, lassen sich an dieser Stelle schon Wort-Embeddings erstellen.

3.1 Datensätze für die Evaluation

Für die Evaluation wurden folgende russische Äquivalente [3] der bekannten Datensätze "Human Judgements of Word Pairs" und "SimLex" verwendet:

Datasets:
• hj + Subsets: a. hj-mc b. hj-rg c. hj-wordsim353-relatedness d. hj-wordsim353-similarity • simlex999

Um das beste Modell zu finden, wurde für alle Modelle der Durchschnittswert des Spearman-Rangkorrelationskoeffizienten von allen Datensätzen ermittelt.

3.2 fastText

Für das in dieser Arbeit erstellte Korpus bietet sich fastText besonders gut an, erstens, weil es für die russische Sprache wichtige Subwortinformationen enthält, und zweitens, wie die Erkenntnisse von Bojanowski, Grave et al. (2016) [4] nahelegen, liefert fastText auch dann gute Ergebnisse, wenn kleinere Korpora zur Verfügung stehen. Die Autoren bieten zusätzlich eine schnelle C++-Implementation mit Python-Bindings (als Python-Modul "fasttext" installierbar), welcher im Rahmen dieser Arbeit der Vorzug gegeben wurde.[1]

Für das Training wurden alle Kombinationen von folgenden Parametern verwendet:

Parameter	Werte
Corpus	original corpus, lemmatized corpus, lemmatized corpus with POS-tags
Model	skipgram, cbow
Number of epochs	10, 20
Dimensions	100, 200
Learning rate	0.05, 0.1
Minimum/maximum length of N-gram	3, 6
Minimum count in vocabulary	2

Die Ergebnisse wurden des Weiteren mit denen des aktuellen fastText-Modells von Facebook fürs Russische verglichen, sowie auch mit den Ergebnissen des Modells von

[1] Die fastText -Implementation von gensim war wegen der Geschwindigkeit (auch mit dem installierten C-Compiler) nicht zufriedenstellend.

RusVectores[1] [5], das auf dem Taiga-Korpus [6] trainiert wurde, und sehen für meine jeweils TOP-4 Modelle so aus:

Dataset: / Models best:	MC	RG	WS353-rel	WS353-sim	HJ	SimLex-999	Average	OOV %
tayga_upos_skipgram_300	0.81	0.82	0.74	0.62	0.8	0.43	0.7	1.35
cc.ru.300[2]	0.76	0.68	0.55	0.67	0.64	0.28	0.6	2.16
Models without POS-Tags (lemmatized):								
ft100_sk_lem_e20_lr.1	0.65	0.6	0.35	0.61	0.52	0.13	0.48	19.22
ft200_sk_lem_e10_lr.1	0.65	0.64	0.26	0.58	0.49	0.15	0.46	
ft100_sk_lem_e20_lr.05	0.63	0.57	0.22	0.59	0.45	0.14	0.44	
ft200_sk_lem_e20_lr.05	0.66	0.57	0.24	0.58	0.46	0.14	0.44	
Models with POS-Tags (lemmatized):								
ft200_sk_lem_pos_e10_lr.1	0.7	0.59	0.25	0.6	0.46	0.12	0.45	17.46
ft100_sk_lem_pos_e20_lr.1	0.63	0.55	0.36	0.62	0.49	0.04	0.45	
ft200_sk_lem_pos_e20_lr.1	0.58	0.6	0.32	0.58	0.48	0.13	0.45	
ft100_sk_lem_pos_e20_lr.05	0.64	0.55	0.31	0.6	0.47	0.11	0.45	
Models without POS-Tags (original):								
ft100_cb_e20_lr.1.bin	0.64	0.62	0.37	0.49	0.48	0.04	0.44	38.57
ft200_sk_e10_lr.1.bin	0.64	0.66	0.29	0.49	0.43	0.08	0.43	
ft100_sk_e10_lr.1.bin	0.65	0.63	0.29	0.47	0.41	0.05	0.42	
ft200_sk_e20_lr.05.bin	0.61	0.6	0.26	0.48	0.41	0.08	0.41	

An dieser Stelle muss darauf hingewiesen werden, dass das Modell von RusVectores POS-Tagging (UPOS) benutzt, deshalb müssen alle Datensätze getaggt werden (die POS-Tags, die MyStem zurückgibt sind mit den UPOS nicht kompatibel, deshalb müssen die in dieses Format konvertiert werden).

Das Strugatzki-Modell reicht noch nicht an die komplexeren und viel größeren Modelle von Facebook oder RusVectores heran und kann mit jenen natürlich auch nicht konkurrieren. Abgesehen von den allgemein relativ schlechten Ergebnissen auf "simlex999", sind dessen Ergebnisse jedoch auf den anderen Datensätzen zumindest nicht sehr schlecht. Die durchschnittliche OOV-Rate, die man auf der originalen Version des Korpus hat, ist inakzeptabel hoch[3], was bedeutet, dass von diesen Modellen in dieser Arbeit kein Gebrauch gemacht wird. Bei den Modellen mit eingeschaltetem POS-Tagging ist sie etwas niedriger als bei denen ohne POS-Tags, was wahrscheinlich darauf zurückzuführen ist, dass die POS-N-Grame dem Modell helfen, bei unbekannten Wörtern trotzdem einen Vektor zu finden, da es kein unbekanntes POS-Tag gibt.

Die Ergebnisse sind dennoch im Großen und Ganzen für den Zweck und die Fragestellungen dieser Arbeit zufriedenstellend.

[1] Das SGNS-Model von RusVectores, das auf dem 5B-großen Webkorpus basiert, das hauptsächlich aus literarischen Werken besteht (aus diesem Modell wurde der Teil mit Poesie ausgeschlossen).
[2] Das fastText-Model von Facebook fürs Russische, das auf Wikipedia-Daten trainiert wurde.
[3] Und ist leicht zu erklären: viele unflektierte Formen sind selten, was der Grund dafür ist, dass man sie im Korpus nie zu sehen bekommt.

3.3 GloVe

Zum Vergleich wurden auch GloVe-Embeddings [7] mit Hilfe des von der Universität Stanford angebotenen Tools trainiert. Folgende Parameter wurden angewendet:

Parameter	Werte
Corpus	lemmatized corpus
Number of epochs	10, 20, 30
Dimensions	100, 200
Window size	10, 15, 20
Minimum count in vocabulary	2
X_{max}	10, 20, 50, 100

Ergebnisse der Evaluation für die TOP-4:

Dataset: / Models best:	MC	RG	WS353-rel	WS353-sim	HJ	SimLex-999	Average	OOV %
glove_300d_30it_15win_xmax10	0.33	0.23	0.23	0.43	0.32	0.02	0.26	19.22
glove_300d_30it_15win_xmax20	0.25	0.24	0.23	0.42	0.33	0	0.25	
glove_300d_30it_15win_xmax50	0.27	0.26	0.17	0.34	0.27	-0.04	0.21	
glove_100d_30it_20win_xmax10	0.24	0.26	0.13	0.26	0.21	-0.06	0.17	

Wie man sehen kann, liefern auch die besten GloVe-Modelle deutlich schlechtere Ergebnisse als die von fastText. Aus diesem Grund werden die Letzteren, namentlich das Modell "ft100_sk_lem_e20_lr.1", im weiteren Verlauf dieser Arbeit bevorzugt. Unerwartet ist des Weiteren die Tatsache, dass die OOV-Rate in den beiden Modellen auf dem gleichen Niveau geblieben ist.

4 Messung von Genderbias

4.1 Einführung

Einer der Tests, den man durchführen kann, um einen möglichen Bias in dem Korpus zu identifizieren, ist in dem Paper von Garg und Kollegen (2018) [8] ausführlich beschrieben. Die Idee besteht darin, eine der Ähnlichkeitsmetriken (wie die negative euklidische Distanz oder Kosinus-Ähnlichkeit) als ein Assoziationsmaß zu verwenden, um feststellen zu können, welche Wörter man eher mit Frauen in Verbindung bringt und welche mit Männern.

Garg und Kollegen haben für diesen Zweck verschiedene Wortlisten erstellt, die sogenannte neutrale Wörter enthielten, die per se nicht genderspezifisch sind, doch sind einige davon in manch einem Modell mit einem Geschlecht mehr assoziiert als mit dem anderen. Ein Gender kann ebenfalls durch eine Wortliste repräsentiert werden, die typische geschlechtsspezifische Wörter inkludiert, wie etwa "Mutter", "Tochter", "Tante", "Frau" usw. Es handelt sich hierbei um einen Durchschnittsvektor von diesen Wörtern.

Nachfolgend ist die Formel angeführt, die in dieser Arbeit benutzt wurde, um den Gender-Bias zu ermessen:

- Für die Assoziation zwischen einem neutralen Wort (v_n) und zwei Gender-Vektoren benutzt man die von Garg et al. definierte "relative norm distance":

$$rnd(v_1, v_2, v_n) = \|v_n - v_1\|_2 - \|v_n - v_2\|_2$$

- Die Autoren haben außerdem angegeben, dass die Wahl der Ähnlichkeitsmetrik unwichtig ist, da die Ergebnisse miteinander signifikant korrelieren (Pearson > 0.95 in den meisten Fällen). Deshalb kann die negative euklidische Distanz auch durch die Kosinus-Ähnlichkeit ersetzt werden[1]:

$$cosinus\ bias(v_1, v_2, v_n) = \cos(v_2, v_n) - \cos(v_1, v_n)$$

Wenn die Zahl positiv ist, dann ist das ein Anzeichen der stärkeren Verknüpfung mit der Gruppe 2 (in der vorliegenden Arbeit gehören Frauen zur Gruppe 2).

4.2 Zusammenstellung der Wortlisten

Wie oben schon angeführt, wurden von Garg und Kollegen verschiedene Wortlisten vorgeschlagen. Im Rahmen meiner Arbeit werde ich auf die jeweiligen Übersetzungen angewiesen sein, die ich mit Hilfe von Google Translate erstellt und anschließend revidiert habe. Das Letztere war deswegen notwendig, da die Qualität der Übersetzung einzelner Wörter meistens nicht so qualitativ ist, obwohl der Übersetzer von Google derzeit sehr gut sprachliche Kontexte erfassen kann. Der zweite Grund ist, dass das Strugatzki-Modell verhältnismäßig klein ist und nicht alle Wörter enthält. Darauf wurde geachtet und jene Wörter, die dem Modell nicht bekannt waren, sind, wenn möglich, durch entsprechende Synonyme ersetzt worden.

[1] Um den Assoziationsgrad mit der ganzen Liste zu berechnen, nimmt man einfach die Summe von allen rnd's für die jeweilige Wortliste: $\sum_{v_n \in N} \cos(v_2, v_n) - \cos(v_1, v_n)$

Im Folgenden werden Wortlisten angeführt, die ich für die Bias-Analyse von Garg et al (2018) übernommen habe[1]:

1. personalitytraits_original.txt – Alle Persönlichkeitsmerkmale (347 Wörter)
2. adjectives_intelligencegeneral.txt – Intelligenz (26 Wörter)
3. adjectives_appearance.txt – Aussehen (23 Wörter)
4. adjectives_negative.txt – Negative Eigenschaften (18 Wörter)
5. female_pairs.txt
6. male_pairs.txt

Was die letzten beiden angeht, so wurden sie für den Zweck dieser Arbeit ein wenig modifiziert. Es wurde auf einige Wörter verzichtet und ein paar neue wurden hinzugefügt. Die Änderungen kann man in der darauffolgenden Tabelle sehen[2]:

Original	Weggelassen	Hinzugefügt
she, daughter, hers, her, mother, woman, girl, herself, female, sister, daughters, mothers, women, girls, females, sisters, aunt, aunts, niece, nieces	• **Alle Pluralformen**, weil das Korpus nur Singularformen enthält • **herself, her, hers**, weil sie zum gleichen Lemma führen wie ihre maskulinen Äquivalente	• Госпожа (господин)[3] • Девушка (парень)[4] • **Optional**: TOP-10 Namen

Da es sich im Falle der Werke, die für das in dieser Arbeit erstellte Korpus verwendet wurden, um Prosadichtung handelt, wäre es von Interesse, außer den üblichen das Geschlecht repräsentierenden Wörtern, für einen der Bias-Tests auch die häufigsten Namen der Hauptcharaktere mit einzubeziehen[5].

4.3 Assoziierte Wörter– Strugatzki

Im Folgenden werden die TOP-10 mit dem jeweiligen Geschlecht assoziierten Persönlichkeitsmerkmale aufgeführt. Die Gender-Vektoren (die Definitionen entsprechen den Tests, die durchgeführt wurden) waren wie folgt definiert:

1. Modifizierte Gender-Listen von Garg und Kollegen ohne Verwendung von Eigennamen der Hauptcharaktere
2. Dieselben Listen wie in 1), aber durch 10 Eigennamen jeweils ergänzt
3. Nur Eigennamen

Das Problem, das mit der Benutzung der Eigennamen auftritt, ist, dass die männlichen Namen sehr viel häufiger sind (1815 vs. 200 für den jeweils häufigsten), deshalb kann man vermuten, dass auch die meisten Persönlichkeitsmerkmale mit ihnen kontextuell häufiger

[1] Sind auf der Github-Seite des Projektes verfügbar: https://github.com/nikhgarg/EmbeddingDynamicStereotypes/tree/master/data
[2] In der Tabelle sind nur weibliche Formen aufgelistet, da sie zu den weiblichen komplett äquivalent sind. Die männlichen Formen wurden auf dieselbe Weise verändert.
[3] Anredeform von "Frau" (bzw. "Herr").
[4] = Junge Frau (junger Mann).
[5] Die Namen kann man unter Einsatz von MyStem aus dem Korpus extrahieren und ihre Vorkommenshäufigkeit zählen.

auftreten würden. Aus diesem Grund habe ich nur die Namen ausgewählt, die für die beiden Geschlechter vergleichbar oft im Korpus vorliegen.

Die Ergebnisse dieser Tests lassen sich den untenstehenden Tabellen entnehmen:

Frauen		Männer	
Original	**Übersetzung**	**Original**	**Übersetzung**
neurotic	нервный	malicious	злобный
gentle	нежный	desperate	отчаянный
elegant	элегантный	presumptuous	самонадеянный
open	открытый	proud	гордый
confused	смущенный	moody	угрюмый
disturbing	тревожный	ungrateful	неблагодарный
trusting	доверчивый	willful	упрямый
loyal	лояльный	aggressive	агрессивный
erratic	непредсказуемый	helpful	полезный
retiring	застенчивый	strong	сильный

Tabelle 1. TEST 1: TOP-10 Persönlichkeitsmerkmale, ohne Verwendung von Eigennamen

Frauen		Männer	
Original	**Übersetzung**	**Original**	**Übersetzung**
elegant	элегантный	predatory	хищный
gentle	нежный	aggressive	агрессивный
trusting	доверчивый	greedy	жадный
open	открытый	malicious	злобный
vulnerable	уязвимый	daring	отважный
tough	жесткий	desperate	отчаянный
neurotic	нервный	willful	упрямый
impersonal	беспристрастный	proud	гордый
sweet	милый	tasteless	пошлый
orderly	опрятный	driving	движущий

Tabelle 2. TEST 2: TOP-10 Persönlichkeitsmerkmale, mit Verwendung von Eigennamen

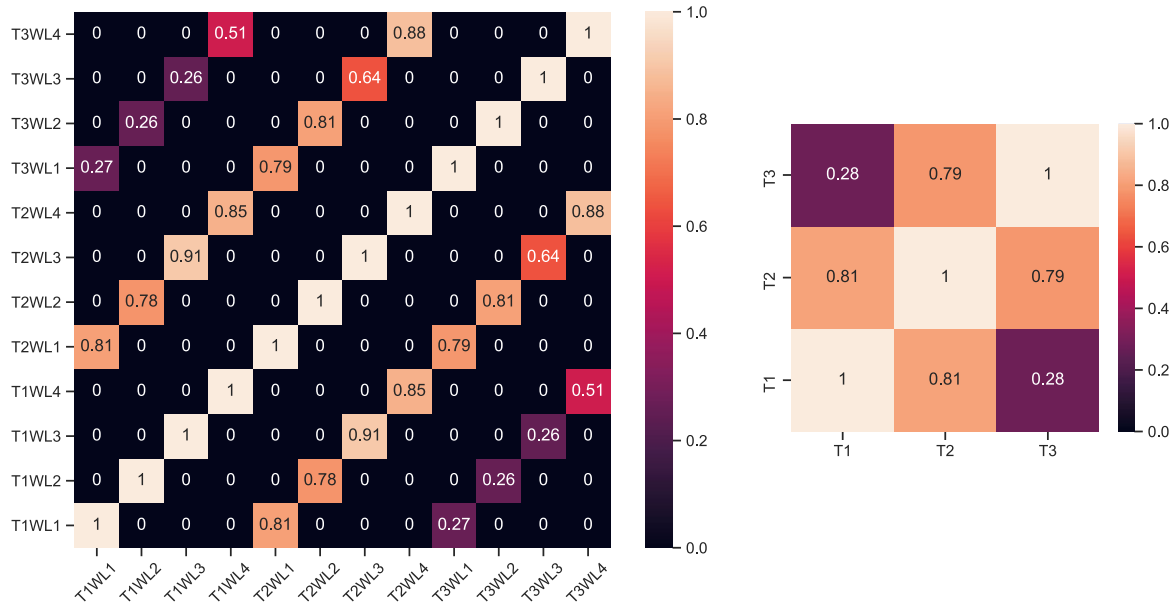
Frauen		Männer	
Original	**Übersetzung**	**Original**	**Übersetzung**
tough	жесткий	tasteless	пошлый
orderly	опрятный	physical	физический
vulnerable	уязвимый	greedy	жадный
dogmatic	догматик	predatory	хищный
trusting	доверчивый	scholarly	ученый
honorable	почтенный	kind	добрый
solitary	уединенный	dynamic	активный
mature	зрелый	daring	отважный
abrupt	грубый	subjective	субъективный
warm	теплый	superficial	поверхностный

Tabelle 3. TEST 3: TOP-10 Persönlichkeitsmerkmale, nur Eigennamen

Wenn man nur von diesem Ranking ausgehen würde, könnte man fast behaupten, dass es sich hierbei um einen spürbaren Bias den Männern gegenüber handelt. Gerade mit ihnen sind die uncharmantsten Qualitäten stark assoziiert (wie "tasteless", "predatory", "greedy" usw.). Was die Frauen-TOP-10 anbelangt, so zeugen hauptsächlich die ersten beiden Tests

von einer stereotypen Darstellung der für Frauen als kennzeichnend wahrgenommenen Eigenschaften ("gentle", "elegant", "trusting", "sweet").

Den Pearson-Korrelationskoeffizienten zwischen den Bias-Scores einzelner Wortlisten in den drei Tests sieht man anhand folgender Heatmaps[1]:



Aus den vorliegenden Heatmaps geht hervor, dass der Zusatz von Eigennamen die Resultate markant verändert: die beiden ersten Tests korrelieren noch sehr stark miteinander (da die das Gender repräsentierenden Wörter aus T1 eine Untermenge von denen aus T2 sind). Wenn man jedoch die Gender-Wörter, auf die man sich in [8] verließ, komplett weglässt, bekommt man ganz andere Ergebnisse. Außer der Wortliste mit allen intelligenzbezogenen Wörtern, die am meisten durch die unterschiedlichen Definitionen betroffen wurde (Korrelation im Durchschnitt = 0.52), kann man die Auswirkungen auch an den anderen Wortlisten erkennen.

Die Bias-Scores (hier sowie im weiteren Verlauf der Arbeit wird der durchschnittliche Wert angegeben, damit die Ergebnisse miteinander vergleichbar sind) für die vollen Listen deuten in allen Fällen auf eine stärkere Assoziation mit Männern hin und sehen wie folgt aus (neben dem Bias-Score ist in der Tabelle auch die Prozentzahl der Wörter mit dem negativen Vorzeichen des Bias-Wertes angegeben):

Test \ WL	1 – personality traits	2 – intelligence	3 – appearance	4 – negative
1	-0.032 (71%)	-0.036 (73%)	-0.019 (56%)	-0.042 (83%)
2	-0.032 (75%)	-0.035 (69%)	-0.019 (56%)	-0.027 (56%)
3	-0.027 (68%)	-0.029 (73%)	-0.015 (61%)	-0.01 (55%)

Tabelle 4. Bias-Scores für einzelne Wortlisten mit der Prozentzahl von Wörtern, die stärker mit Männern assoziiert werden

[1] Einmal für einzelne Wortlisten und einmal für Tests insgesamt. "T" steht für "Test", "WL" für "Wortliste". Die Reihenfolge der Wortlisten korrespondiert mit der aus dem Abschnitt 4.2

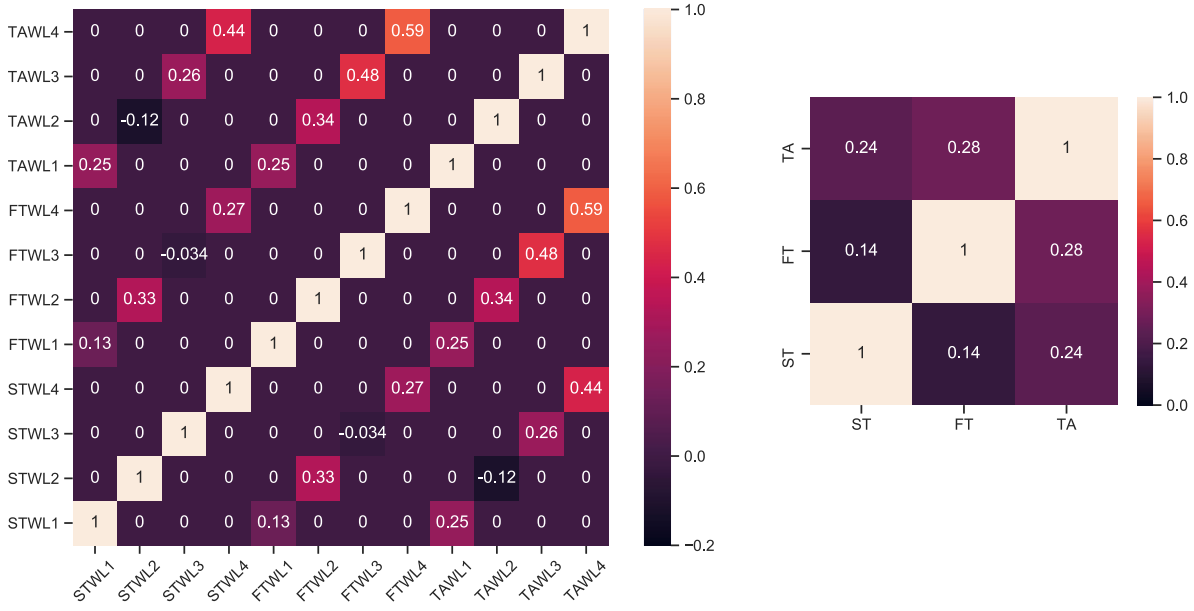
Man kann konstatieren, dass die Ergebnisse bedeutend davon abhängen, was für Definitionen man für den das Geschlecht repräsentierenden Vektor verwendet. Das bestätigt des Weiteren auch die Möglichkeit, dass die Erkenntnisse von Garg et al. (2018) viel anders ausgesehen hätten, hätten die Autoren bspw. für die Ethnien etwas anderes als ihre Nachnamen benutzt.

5 Vergleich mit den aktuellen Modellen

Das Ziel dieses Abschnittes ist, die Bias-Werte des Strugatzki-Modells mit denen von aktuell zwei besten Modellen[1] für die russische Sprache zu vergleichen:

1. Mit dem fastText-Modell, die die gegenwärtige Sprachverwendung, wie sie in verschieden Artikeln dokumentiert ist, gut darstellen soll
2. Mit dem Taiga-Modell, das für die literarische Sprache mit all ihren Vorurteilen repräsentativ sein soll

Um diese drei Modelle miteinander vergleichen zu können, habe ich die Gender-Definition aus dem ersten Test des Strugatzki-Modells übernommen (denn es ergäbe relativ wenig Sinn, die Namen der Hauptfiguren der Strugatzki-Werke in die Analyse der zwei davon unabhängigen Modelle mit einzubeziehen). Die Korrelation kann man den unten aufgeführten Heatmaps entnehmen:



Es lässt sich festhalten, dass die Ergebnisse dieser drei Modelle sehr unterschiedlich sind. Das Strugatzki-Modell zeigt dennoch eine stärkere Korrelation mit dem Taiga-Modell. Interessant ist des Weiteren, dass man den höchsten Korrelationswert zwischen der Wortliste mit negativen Eigenschaften (WL4) des fastText- und Taiga-Modells sieht, beide scheinen das Negative eher mit Männern zu verknüpfen.

TOP-10 Persönlichkeitsmerkmale, die aus diesen beiden Modellen hervorgehen, sehen wie folgt aus[2]:

[1] Erwähnt in Abschnitt 3.2

[2] Bei dem fastText-Modell war es nicht möglich, die TOP-10 anzugeben, da nur 5 Persönlichkeitsmerkmale mehr mit Frauen in Verbindung gesetzt werden als mit Männern. Das betrifft darüber hinaus auch die anderen Wortlisten – ein paar Ausnahmen ausgenommen, war der Bias-Wert immer im negativen Bereich

Frauen		Männer	
Original	**Übersetzung**	**Original**	**Übersetzung**
whimsical	капризный	treacherous	предатель
crisp	хрупкий	stoic	мужественный
gentle	нежный	cowardly	трусливый
fickle	непостоянный	moody	угрюмый
hysterical	истеричный	rustic	простоватый
sensual	чувственный	deceptive	лжец
cooperative	приветливый	paternal	отеческий
irrational	нерациональный	ridiculous	чудаковатый
repressed	скованный	idealistic	идеалист
warm	теплый	scholarly	ученый

Tabelle 1. TOP-10 Persönlichkeitsmerkmale – Taiga

Frauen		Männer	
Original	**Übersetzung**	**Original**	**Übersetzung**
troublesome	проблемный	ambitious	честолюбивый
earnest	серьезный	ridiculous	чудаковатый
ruined	испорченный	brilliant	гениальный
aspiring	стремящийся	sly	хитрый
knowledge	знание	dignified	благородный
		boisterous	неистовый
		colorless	неприметный
		crafty	коварный
		willful	упрямый
		charming	обаятельный

Tabelle 2. TOP-10 (bzw. 5) Persönlichkeitsmerkmale – fastText

Man kann ausgehend von den Ergebnissen des Taiga-Modells (im Gegensatz zum fastText-Modell) ein ähnliches Muster erkennen, welches auch die Autoren von [8] in ihren Modellen feststellen konnten, in welchen ein Bias zu finden war.

WL / Modell	1 – personality traits	2 – intelligence	3 – appearance	4 – negative
Taiga	-0.019 (70%)	-0.047 (96%)	-0.001 (48%)	-0.017 (72%)

Tabelle 3. Bias-Scores für einzelne Wortlisten mit der Prozentzahl von Wörtern, die stärker mit Männern assoziiert werden

Obwohl alle Zahlen negativ sind, ist zum Beispiel der absolute Durchschnittswert für WL2 um 47mal größer als der für WL3. Es könnte von Interesse sein, die beiden Listen einander gegenüberzustellen (mit den jeweils 10 größten Bias-Werten):

Intellekt		Aussehen	
Word	**Bias**	**Word**	**Bias**
intuitive	0.019	pretty	0.146
adaptable	-0.008	alluring	0.101
brilliant	-0.012	sensual	0.071
ingenious	-0.023	beautiful	0.061
imaginative	-0.025	homely	0.056
logical	-0.029	attractive	0.048
intelligent	-0.029	slender	0.043
reflective	-0.031	slim	0.035
astute	-0.032	fashionable	0.023
analytical	-0.034	plump	0.02

Anhand der beiden letzten Tabellen wird es deutlich, dass der Bias-Wert allein (bzw. sein Vorzeichen) noch nicht aussagekräftig ist, daher ist es unbedingt erforderlich ihn mit den anderen Werten zu vergleichen. In diesem Fall sind bspw. alle Adjektive, die ein hohes intellektuelles Vermögen beschreiben, stärker mit Männern verknüpft und die, die jemandes Anmut schildern, sind für Frauen charakteristisch.

6 Fazit

Die Ergebnisse haben gezeigt, dass ein Bias in den Werken der Brüder Strugatzki tatsächlich festzustellen ist, was sich in der Art von Assoziationen widerspiegelt, und zwar waren die beiden Geschlechter am besten durch relativ stereotyphafte Bezeichnungen dargestellt. Allerdings ist die Anzahl der Methoden, um den Bias in solch einem Korpus zu quantifizieren, sehr bescheiden. In diesem Fall verließ man sich nur auf die Assoziationstests, was womöglich unzureichend ist, um die Werke anhand derer Ergebnisse für vorurteilsvoll zu deklarieren. Das wirkt sich auch auf die Validierung dieser Ergebnisse aus, da es am Ende nicht so viel gibt, was man validieren kann. Nichtsdestotrotz erfasst mein Modell zumindest die wichtigsten Assoziationen verhältnismäßig gut, was in Anbetracht seiner Größe bemerkenswert ist.

Garg und Kollegen standen statistische Daten zur Verfügung, welche einem Einblick geben könnten, wie sich die Vorurteile im Laufe der Zeit veränderten. Es war gewünscht, derartiges auch auf das in dieser Arbeit erstellte Korpus anwendbar machen zu können, wie beispielsweise einzelne Modelle für die jeweiligen Schaffensphasen der Brüder Strugatzki zu kreieren (für frühes, mittleres und spätes Werk z.B.). Dieses Vorhaben erschien aber angesichts der Korpusgröße eher unrealistisch – das Modell ist zu klein dafür, was dazu führen könnte, dass die Ergebnisse nicht konklusiv wären, vielmehr willkürlich.

Was auch zu beobachten war, ist die Tatsache, dass eine höhere Korrelation des von mir erstellten Modells mit den Ergebnissen des Modells zu verzeichnen war, das an literarischen Werken trainiert wurde, also mit dem Modell, das sich (aus inhärenten Gründen) sehr "gefärbt" gezeigt hat.

Obwohl die Ergebnisse im Falle eines kleinen Korpus von Werken zweier Autoren ziemlich schwer zu interpretieren sind, wäre es dennoch möglich, fast alle von Garg et al. (2018) erwähnten Methoden auf das viel größere Korpus aller literarischen Werke anzuwenden, vorausgesetzt, es lassen sich gute Submodelle davon trainieren.

7 References

[1] A. Strugatsky and B. Strugatsky, "Books. Complete set of works," 1998-2015. [Online]. Available: http://www.rusf.ru/abs/. [Accessed March 2020].

[2] I. Segalovich, „A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine.," 2003.

[3] A. Panchenko, D. Ustalov, N. Arefyev, D. Paperno, N. Konstantinova, N. V. Loukachevitch und C. Biemann, „Human and Machine Judgements for Russian Semantic Relatedness," *CoRR,* Bd. abs/1708.09702, 2017.

[4] P. Bojanowski, E. Grave, A. Joulin und T. Mikolov, „Enriching Word Vectors with Subword Information," *CoRR,* Bd. abs/1607.04606, 2016.

[5] A. Kutuzov und E. Kuzmenko, „WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models," 2017.

[6] T. Shavrina and O. Shapovalova, "To the Methodology of Corpus Construction for Machine Learning: „Taiga" Syntax Tree Corpus and Parser," in *Proceedings of the International Conference „Corpus Linguistics 2017"*, St. Petersburg, 2017.

[7] J. Pennington, R. Socher und C. D. Manning, „GloVe: Global Vectors for Word Representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[8] N. Garg, L. Schiebinger, D. Jurafsky und J. Zou, „Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes," *CoRR,* Bd. abs/1711.08412, 2017.