Exercise 2.1: N-grams Frequency Analysis

In the following part of work we collected n-gram frequency counts for German and English text corpora.

a) Word tokens of each text corpus were converted into a set of n-grams (up to n = 4). For this purpose a method *char_ngrams* was created, it takes as an input a word token and and an integer m and returns a list of all possible n-grams in the token from n = 1 up to n = m.

b) Then we generated a table of the top 15 frequent character unigrams (1-grams), bigrams (2-grams), trigrams (3-grams), and 4-grams for each text corpus. The most common n-grams of text corpora are represented in Table 1 for English and German.

Table 1. The top 15 frequent character n-grams in English and German.

	English				German			
	Uni-grams	Bi-grams	Tri-grams	Four-grams	Uni-grams	Bi-grams	Tri-grams	Four-grams
1	e	th	the	tion	e	en	der	sche
2	t	he	ion	atio	n	er	ung	chen
3	i	in	tio	ment	r	de	sch	isch
4	o	on	ing	sion	i	ch	die	eine
5	n	ti	and	emen	s	un	ich	lich
6	a	an	ati	comm	t	te	che	icht
7	r	re	ent	arti	a	ei	ein	rung
8	s	er	men	ions	d	ie	gen	ngen
9	c	at	for	with	u	ge	und	iche

10	h	io	pro	hall	g	in	den	ungs
11	l	of	ate	shal	h	ng	ten	unge
12	d	or	com	rtic	l	es	ver	über
13	u	en	con	ting	o	nd	nde	komm
14	m	nt	ter	that	m	be	hen	ende
15	p	es	sio	ther	c	st	cht	nder

c) Assuming that the identity of the languages and the original word forms in the two corpora are unknown, it is possible to identify the language based on the distribution of frequent n-grams. According to the Zipf’s law words that are most frequently used tend to be shorter. In this case we could suppose that some of the n-grams from the top-list are represented by words, most frequently used in English or German. Furthermore, other n-grams from the list can be recognized as prefixes and suffixes, which are characteristic for the languages.