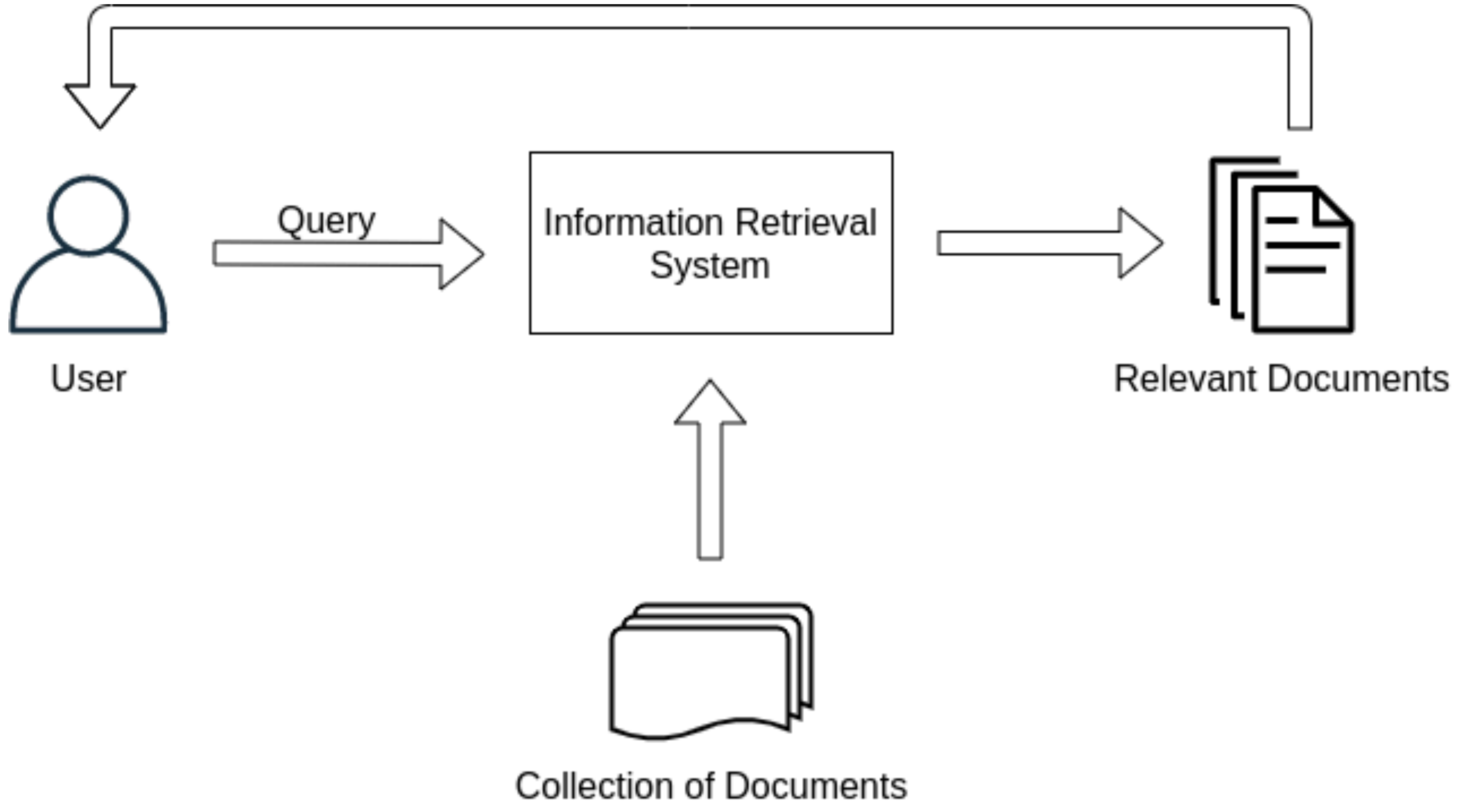
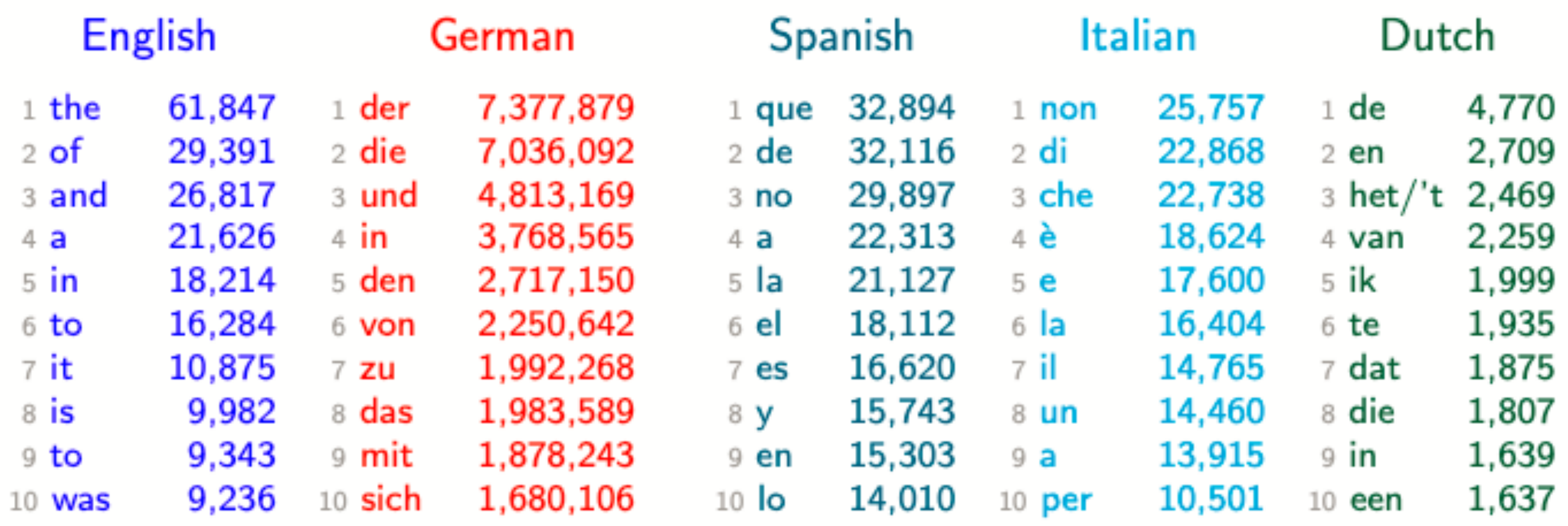
Information retrieval

- Information retrieval refers to a set of methods by which text documents and other unstructured forms of data are matched against some user input query, which is typically text, based on their semantic relevance to that query.
- Examples:
  - Web search
  - Filesystem search
  - Legal and medical record document search
  - Recommender systems can (essentially) be thought of as contextual information retrieval systems



Zipf’s Law and informativeness

- Remember Zipf’s Law? It describes the distribution of word frequencies found in natural text.
- A key feature that we wish to endow IR systems with is the ability to focus on informative words that carry meaning, and ignore ones that don’t. Empirically we know that high frequency words carry far less meaning that those words on the long-tail of the word frequency distribution.
- Given this, a reasonable goal is to build IR systems that are sensitive to low frequency words, and insensitive to high frequency ones.



	English			German			Spanish			Italian			Dutch	
1	the	61,847	1	der	7,377,879	1	que	32,894	1	non	25,757	1	de	4,770
2	of	29,391	2	die	7,036,092	2	de	32,116	2	di	22,868	2	en	2,709
3	and	26,817	3	und	4,813,169	3	no	29,897	3	che	22,738	3	het/'t	2,469
4	a	21,626	4	in	3,768,565	4	a	22,313	4	è	18,624	4	van	2,259
5	in	18,214	5	den	2,717,150	5	la	21,127	5	e	17,600	5	ik	1,999
6	to	16,284	6	von	2,250,642	6	el	18,112	6	la	16,404	6	te	1,935
7	it	10,875	7	zu	1,992,268	7	es	16,620	7	il	14,765	7	dat	1,875
8	is	9,982	8	das	1,983,589	8	y	15,743	8	un	14,460	8	die	1,807
9	to	9,343	9	mit	1,878,243	9	en	15,303	9	a	13,915	9	in	1,639
10	was	9,236	10	sich	1,680,106	10	lo	14,010	10	per	10,501	10	een	1,637

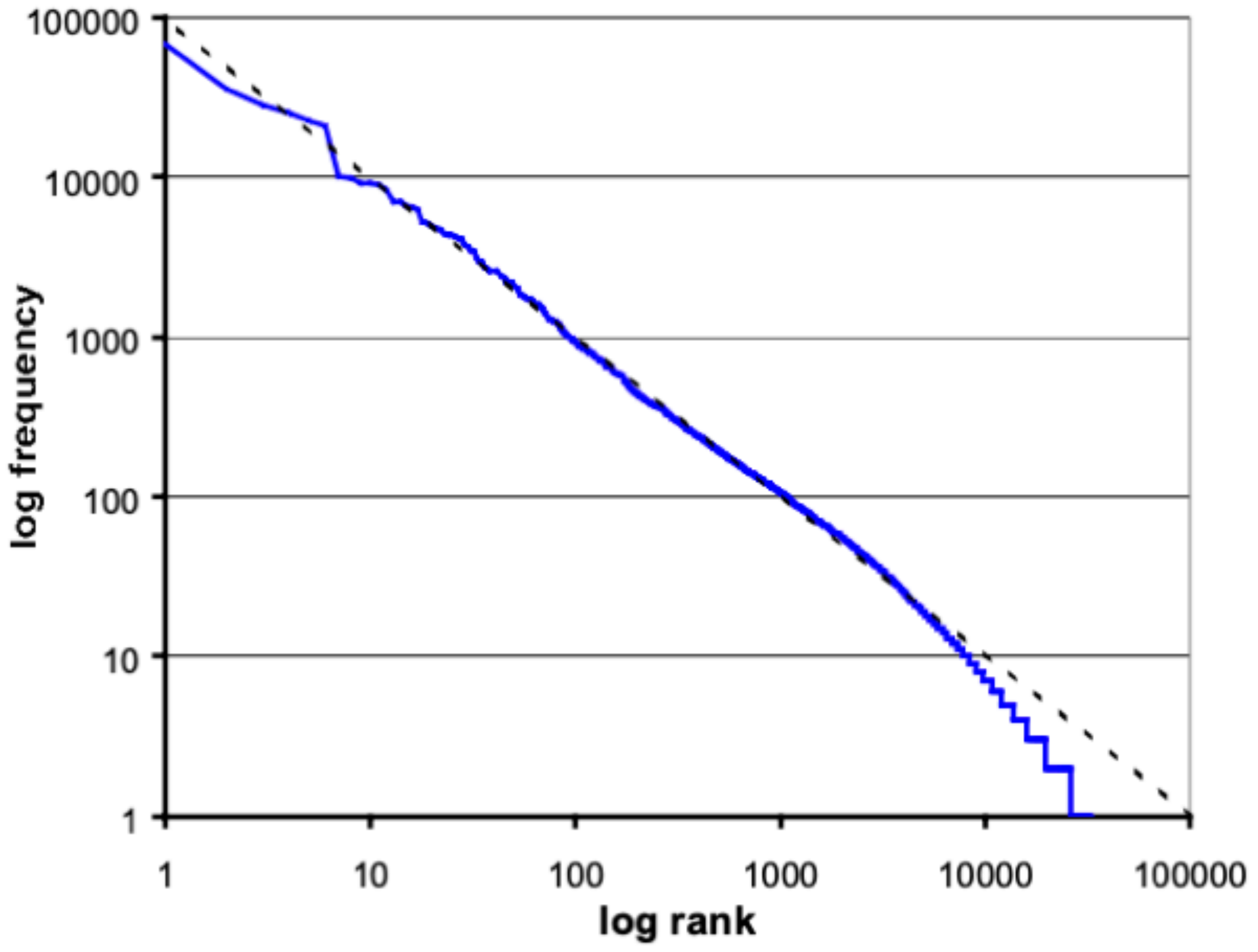


Figure 2 Zipf curve for the unigrams extracted from the 1 million words of the Brown corpus

-Ha et al., *Extension of Zipf’s Law to Words and Phrases*, 2002