

THE DIGITAL DICTIONARY OF THE GERMAN LANGUAGE (DWDS)

BERLIN-BRANDENBURG ACADEMY OF SCIENCES AND HUMANITIES

DWDS (WWW.DWDS.DE)

...is a lexical information system developed by and hosted at the BBAW. The system offers one-click-access to three different types of resources: a) Lexical resources such as the *DWDS-Wörterbuch*, the *Wörterbuch der deutschen Gegenwartssprache* (WDG), *Etymologisches Wörterbuch* by Wolfgang Pfeifer, and the *Deutsches Wörterbuch* by Jacob Grimm and Wilhelm Grimm, b) a large, linguistically annotated German corpus of 20th and 21st century texts containing about 1.8 billion tokens. The DWDS corpus consists of two parts: a core corpus and an extended corpus. The core corpus contains approximately 100 million running words, balanced chronologically and by text genre in approximately 80,000 documents. The extended corpus contains more than 1.7 billion text words. It is an opportunistic corpus, consisting



Figure 1: Web frontend to the DWDS system (www.dwds.de). The frontend aligns several resources in panels which display search results in the respective resource for the search expression ('Troll', in our example).

essentially of major newspaper sources from the last 20 years such as Die ZEIT, Bild, Süddeutsche Zeitung, and Die WELT. Copyright clearance has been obtained from major publishing houses, enabling DWDS

users to access the works of important literary and scientific authors including Heinrich Böll, Jürgen Habermas, Victor Klemperer, Siegfried Lenz, Thomas and Heinrich Mann and Kurt Tucholsky, and also early 20th century sources such as Berliner Tageblatt and Vossische Zeitung; c) statistical resources which are based on the aforementioned corpora, both for individual words (*Wortverlaufskurve*, see figure 2)

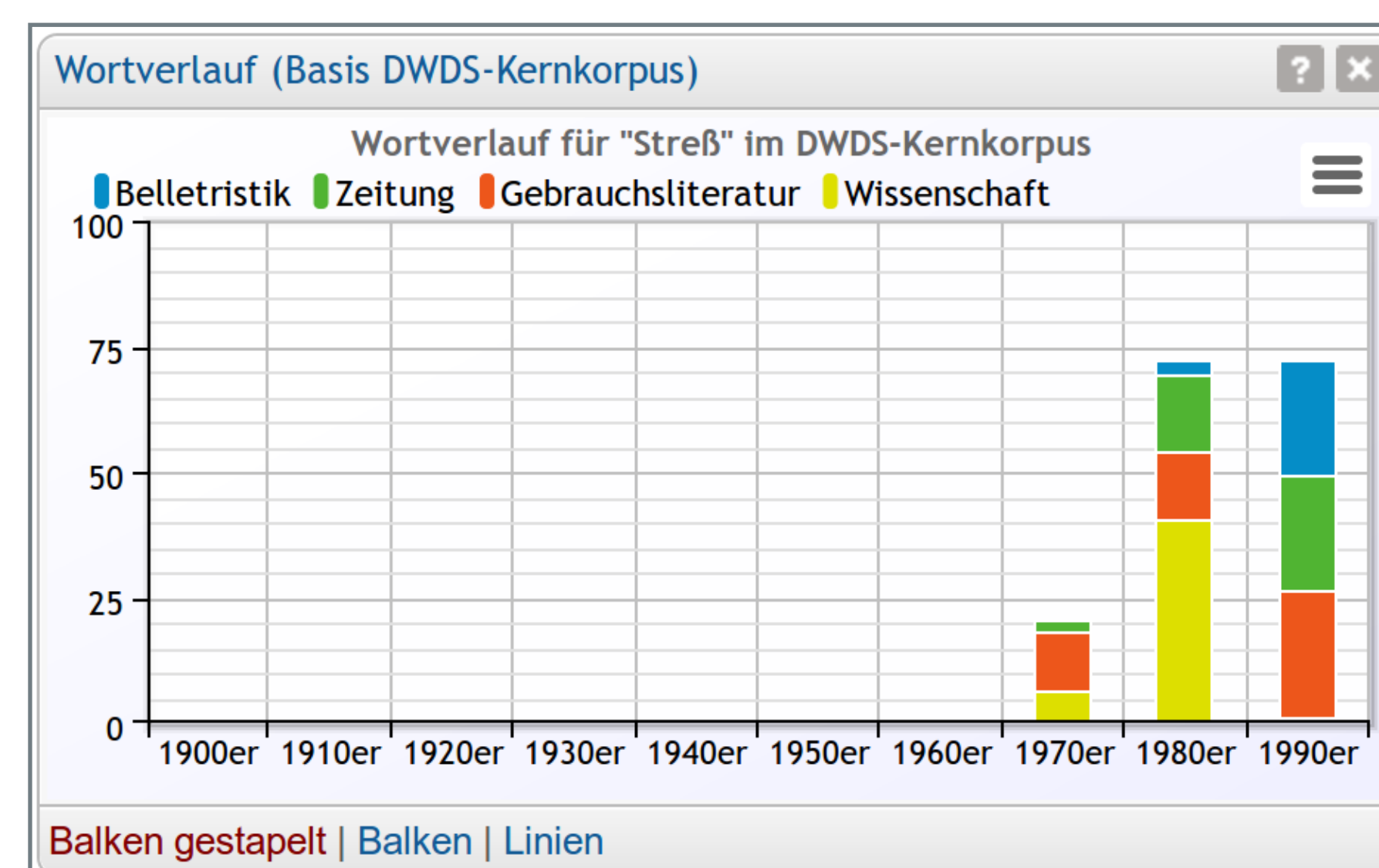


Figure 2: The curve showing the word's history gives a clear picture of its distribution in use across the decades of the 20th century and across different kinds of texts. The evaluation and presentation is based on the DWDS's core corpus of the 20th century. The word 'Streib' (stress) has been in use in German only since the 1970s and at first it appeared primarily in scholarly and functional literature.

and word combinations (*Wortprofil*, see figure. 3). These resources are displayed alongside one another in separate panels (see figure 1). The system offers the choice among several views, i.e. between several profiles with predefined panel combinations.

KEEPING THE RESOURCES UP-TO-DATE

A team of six lexicographers and four resource engineers are currently working on keeping the resources, both dictionaries and corpora, up-to-date. In a complex process of analysing several corpora, 45,000 words have been identified as candidates for inclusion into the *DWDS-Wörterbuch* (among them

widely used words such as 'Handy' and 'Frischkäse'). Additionally, there are numerous words which have acquired new meanings (e.g. the sense of 'abklingen' which is translated as 'to decay' in the context of radioactivity).

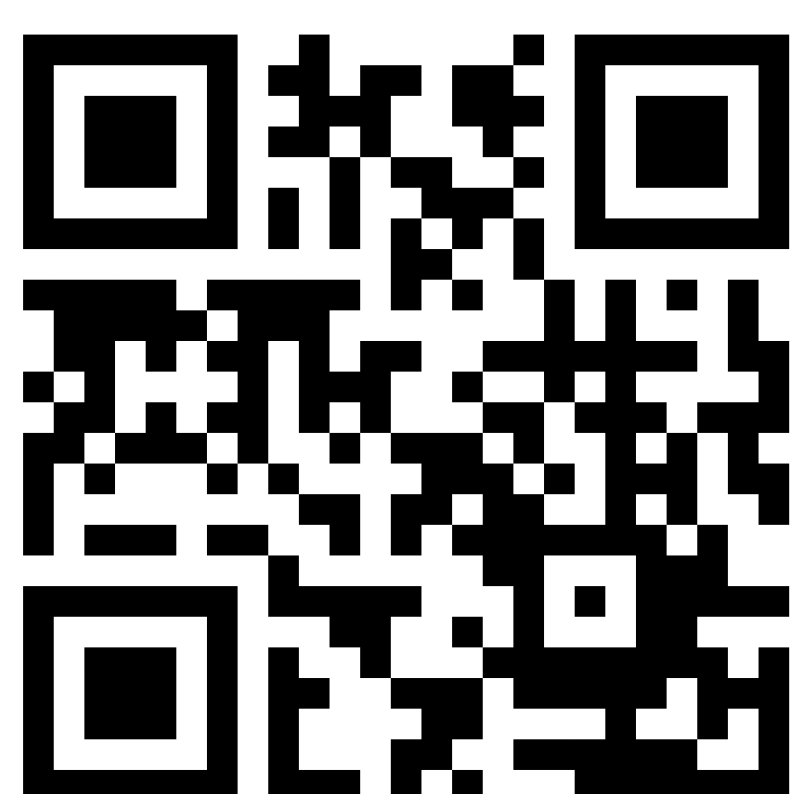
The second type of resource which has to be kept up-to-date is our corpora. We are permanently looking for donors of texts (e.g. newspapers and publishing houses) to stock up our corpus base.

CO-OPERATIONS

Additionally, we are partner in a project which aims at building a large German reference corpus of computer-mediated communication (DeRiK, *Deutsches Referenzkorpus der internetbasierten Kommunikation*). We are also co-operating with other large resource providers, e.g. the Institut für Deutsche Sprache, in a large European Research Infrastructure project (CLARIN, Common Language Resources and Technology Infrastructure).



Figure 3: The word's profile illustrates in the form of a tag cloud the other words with which it typically occurs. It is also possible to filter for the grammatical relation between the two words. The evaluation and presentation is based on the DWDS's core corpus of the 20th century and the ZEIT-Korpus. Here we are shown the words with which the word 'grau' (grey) typically appears. The words displayed in larger size are the 'most typical' co-occurring words ('Eminenz' (eminence), 'Maus' (mouse), 'Panther' (panther) etc.).



Contact
Dr. Alexander Geyken
geyken@bbaw.de
Jägerstraße 22/23
10117 Berlin
www.dwds.de
dwds_de on twitter