Előszó

A *Frequency Dictionaries* sorozat elsődleges célja, hogy összegyűjtse minél több nyelv gyakorisági adatait, hogy ezáltal a nyelvek még inkább összehasonlíthatóvá váljanak. Számos, ebben a sorozatban vizsgált nyelv esetében ez az első ilyen típusú, átfogó és szilárd empirikus alapokon nyugvó gyűjtés.

Jelen szótár nyomtatott és elektronikus formában került kiadásra. Mindegyik feltünteti a leggyakoribb 1,000 szóalakot gyakoriság sorrendben és a leggyakoribb 10,000 szóalakot ábécérendben. A nyomtatott kiadvány részét képezi továbbá a részletes szótári bevezető is, mely a nyelvi adatokkal kapcsolatos információkat és az adatgyűjtés módszertanát tartalmazza, valamint kiegészül szóstatisztikán alapuló, a betűkre és a szavak szerkezetére vonatkozó nyelvspecifikus statisztikai adatokkal, grafikonokkal.

A mellékelt CD-ROM-on elektronikus könyv formájában a szótár átfogóbb változata található, mely 1,000,000 szóalakot ölel fel ábécérendben, relatív gyakorosági adatokkal. A nyelv szóalakjainak száma természetesen függ a mindenkori korpusz méretétől és összetételétől. Ez a gyakorisági osztály adataival ellátott szólista megtalálható a CD-ROM-on egyszerű szövegfájlként is, ábécérendbe és gyakoriság szerint rendezve is. Ezekből már könnyen generálhatóak további szólisták különféle alkalmazások számára. A nyomtatott szótárban található szóalakok ellenőrzése manuálisan történt, hogy kiköszöböljük a hibás vagy elavult alakok felvételét a szótárba, míg a CD-ROM-on található hosszabb listát csak automatikus valószínűségi szempontok figyelembevételével ellenőriztük.

A szótár kizárólag az elektronikusan hozzáférhető, nagyméretű *Leipzig Corpora Collection* adataira támaszkodik. Az egyes gyakorisági szótárak alapját képező korpuszok újságnyelvi szövegeket, Wikipedia szócikkeket és más, az internetről véletlenszerűen gyűjtött szövegeket tartalmaznak. Ezek online is elérhetőek a következő cím alatt: *http://wortschatz.uni-leipzig.de/ws_hun/*.

A gyakorisági szótár-sorozat lehetőséget nyújt összehasonlító vagy csak egyetlen nyelvet érintő nyelvészeti kérdések átfogóbb vizsgálatára, mint amilyen a szóképzés vagy a szókészlet frekvencia-alapú elemzése, pl. szótárak készítése vagy nyelvtanítás céljából. Ezen felül a szótár statisztikai adatai ösztönzően hathatnak más kutatási területekre is. A gyakorisági szótárak címe mindig tartalmazza a vizsgált nyelv nevét angolul, valamint szerepel még a borítón a teljes cím az eredeti nyelven és a mindenkori nyelv ISO 639-3 szerinti hárombetűs rövidítése is.

About the Series

The *Frequency Dictionaries* series aims at producing dictionaries with comparable frequency data for a large number of different languages. For many of the languages featured in this collection, this series is the first comprehensive compilation to use a large-scale empirical base.

The dictionaries are available in both print and electronic versions. Each dictionary provides the most frequent 1,000 word forms in order of frequency and the 10,000 most frequent word forms in alphabetical order. They provide an introductory description of the data and the methodological approach used. In addition, language-specific statistical information is provided with regard to letters, word structure and structural changes.

The enclosed CD-ROM contains a more comprehensive version of the dictionary as an e-book. This includes data on the relative frequency of up to 1,000,000 word forms presented in alphabetical order. The number of word forms for a particular language depends on the size and composition of the corpus used. This list of words (with frequency classes) is also available as a plain text file on the CD-ROM and is ordered both alphabetically and by frequency. Using this file, word lists for various applications can be generated easily. The word forms in the printed part of the dictionary have been checked carefully by hand to identify incorrect forms. In contrast, the more comprehensive list on the CD-ROM has been inspected by means of automatic plausibility criteria alone.

For the compilation, comprehensive electronically available sources of the *Leipzig Corpora Collection* were used consistently. The corpora on which the individual frequency dictionaries are based include newspaper texts, Wikipedia articles and other randomly collected texts available on the Internet. They can be accessed online at *http://wortschatz.uni-leipzig.de/ws_hun/*. This series of dictionaries provides the opportunity to explore comparative linguistic topics and such monolingual issues as studies on word formation and frequency-based examinations of lexical areas for use in dictionaries or language teaching. The statistical results presented here can offer initial suggestions for several areas of research. The title of the frequency dictionaries always includes the name of the language in English, in the original language and using its three-letter abbreviation according to ISO 639-3.

Acknowledgements

The editors and authors would like to thank the members of the Natural Language Processing Group at the Institute of Computer Science at University Leipzig for their invaluable assistance.

Terms of Use

Authors

Dirk Goldhahn: Natural Language Processing Group, Institute of Computer Science at University Leipzig, dgoldhahn@informatik.uni-leipzig.de
Zita Hollós: Károli Gáspár Református Egyetem (Budapest), hollos.zita@kre.hu
Uwe Quasthoff: Natural Language Processing Group, Institute of Computer Science at University Leipzig, quasthoff@informatik.uni-leipzig.de

The frequency dictionaries in this series

Published so far:
Vol.1: GER German - Deutsch
Vol.2: ENG English
Vol.3: ISL Icelandic - Íslenska
Vol.4: FRA French - Français
Vol.5: HUN Hungarian - Magyar

Appearing soon:
EPO Esperanto - Esperanto
NLD Dutch - Nederlands
UKR Ukrainian - Українська
IND Indonesian - Bahasa Indonesia
DAN Danish - Dansk
CES Czech - Čeština

Contents

1 Word forms and their frequencies **1**
1.1 Previous frequency dictionaries for Hungarian 1
1.2 Structure of entries and frequency data 4
1.3 Data in electronic form 5
1.4 Conception of a dictionary of word forms 6
1.5 On the definition of a word 8

2 Corpus **11**
2.1 Data basis and extent of data collection 11
2.2 Corpus timeline 12
2.3 Data pre-processing 13

3 Statistical data on the word list **15**
3.1 Comparison criteria 15
3.2 Character statistics 16
3.3 Vowels and consonants 18
3.4 Word length 19
3.5 Word structure 20
3.6 Zipf's law 23
3.7 Text coverage of the top-N most frequent words 23
3.8 Growth in word length 24
3.9 Number of syllables 27
3.10 Number of letter bigrams and trigrams 27
3.11 Summary of statistical data 29

4 References **31**

5 Most frequent words orderd by rank **33**

6 Alphabetical frequency list **45**