THE ROLE OF CORPORA IN THE WORK OF THE LANGUAGE CONSULTING SERVICE OF THE HUN-REN HUNGARIAN RESEARCH CENTRE FOR LINGUISTICS

Zsófia Ludányi
HUN-REN Hungarian Research Centre for Linguistics, HU
Pázmány Péter Catholic University, HU
ludanyi.zsofia@nytud.hun-ren.hu

Abstract. Relying on Language Management Theory, the paper presents the use of corpus data in language consulting through the work of the Language Consulting Service (LCS) of the HUN-REN Hungarian Research Centre for Linguistics. In the development of language advice, it is essential to take into account the linguistic background, to apply a data-driven approach and, in some cases, to provide access to linguistic data related to specific language problems. The research is based on the LCS database of approximately 10,000 emails, and examines questions whose answering by language consultants was informed by a corpus-based investigation (124 inquiries). The results show that the use of corpus query interfaces is well-suited for answering complex linguistic questions and for describing the usage of particular expressions. On the other hand, corpus data are rarely needed for answering questions on orthography, which constitute the bulk of inquiries that the LCS receives. In answering spelling questions, the basic aim is to communicate the spelling norm, thus the examination of usage is only necessary when spelling codification is absent or inconsistent. The paper presents a case study on a specific language use issue from the database, highlighting in detail the strategy used in the development of language advice, with a particular focus on the appropriate use of corpus query interfaces.

Keywords: language consulting, corpora, corpus querying, language use problems, spelling problems

DOI: https://doi.org/10.12697/jeful.2024.15.2.03

1. Introduction

When performing professional work on texts, such as translation and proofreading, language users face a number of complex language problems which require linguistic knowledge. One way of managing language problems is for language users to contact language consulting institutions. The broad availability of language consulting services and the high number of speakers who use them clearly demonstrate that such linguistic meta-activities play an important role in managing language problems (Biere & Hoberg [eds.] 1995; Riegel 2007; Ludányi 2020).

Based on data from the Language Consulting Service (LCS) of the HUN-REN Hungarian Research Centre for Linguistics, the present study seeks to find out which language problems require the use of corpus query interfaces, i.e. for which problems language consultants tend to rely on corpus data to formulate their responses. The research has been motivated by self-reflection: In recent years the staff of the Hungarian Usage and Language Consulting Research Group running the Language Consulting Service has grown, which has given an opportunity to renew the LCS. As part of the renewal, it is important to review and raise awareness of the consulting practices that have been used in the past, often in an intuitive manner. The use of corpora in language consulting is one such practice. Thus, one contribution of the present research is that it allows the previously intuitive use of corpora to be more principled and systematic. The research results reveal important strategies which may support the process of language consulting and make it more reflexive. The strategies explored may also help those who do professional work on texts (e.g. translators, copywriters, marketing specialists), as they are not necessarily familiar with the use of corpora. Based on our experience, non-linguists in Hungary are only familiar with dictionaries and possibly descriptive grammars, but not at all with corpora.

1.1. Theoretical background

The paper adopts Language Management Theory (LMT, Jernudd & Neustupný 1987), developed from Language Planning Theory (Nekvapil 2006), to describe institutional language consulting as an activity aimed at managing the specific language problems of speakers. The LMT approach to institutional language consulting has been discussed in several works (Beneš et al. 2018; Prošek 2020; Lengar Verovnik & Dobrovoljc 2022; Kopecký 2022; Vranjek Ošlak 2023; Domonkosi &

Ludányi 2023, forthc.; Ludányi & Domonkosi 2023), therefore the subsection below only touches upon the most important basic concepts.

1.1.1. Basic concepts

In the present paper, language management is interpreted as broadly as possible to include all activities directed towards language and specific discourses. Accordingly, it includes reflection on linguistic phenomena and the language-shaping activities that follow from such reflection (Ludányi & Domonkosi 2023). LMT distinguishes between simple and organised language management (Jernudd & Neustupný 1987: 76). While in simple language management, speakers manage individual features or aspects of their own or their discourse partner's language "here and now", in a given interaction, the acts of organised language management are trans-interactional and are implemented by one or more powerful institutions, with a greater emphasis on theorising and ideologies than in simple language management (Nekvapil 2012: 167).

The study interprets interactions between language service inquirers and consultants as a language management cycle, following Beneš et al. (2018: 122–123). The cycle begins when everyday speakers note and evaluate certain linguistic phenomena as problematic or otherwise salient (micro-management). Subsequently, they turn to the language consulting service to develop or facilitate an 'adjustment design' (macro-management). Inquirers may then accept and implement the adjustment design, but they may also reject it or take it only partially into account (micro-management) (Nekvapil 2009: 6). This in turn may prompt the need for a new cycle. From the above discussion it follows that institutional language consulting is crucial for bridging the gap between simple and organised language management processes, i.e. between the micro and macro levels of management (Kimura & Fairbrother 2020).

The paper considers as language problems those linguistic phenomena which are evaluated negatively, positively (Neustupný 2003: 127) or neutrally (Nekvapil 2009: 8),[1] and which elicit some kind of reflection

[1] Neustupný calls *gratification* those linguistic phenomena and situations that are positively evaluated by the speakers. A good example for this from the LCS database is when an inquirer calls the lexeme *illóolaj* 'essential oil' a "beautiful Hungarian word"

from the language user, the speech partner or even a third person observing the interaction from outside (Ludányi & Domonkosi 2023). Such a broad interpretation of the concept is also necessary because inquirers do not only consult the LCS about linguistic phenomena they consider problematic. Rather, they also ask for information about linguistic forms and situations that they note and consider interesting or striking, without expecting guidance or advice.

Based on Lanstyák (2014: 327–328), the paper distinguishes between particular instances of a problem (problem-tokens), which arise in concrete situations, are related to place and time and to concrete people, and the abstract category of so-called meta-problems, which result from the generalisation of problem-tokens. The study considers problem solving to be only one of a variety of ways of managing problems, alongside other involvement strategies such as devolution, alleviation, ignoring, mitigation, acceptance, and non-involvement strategies such as avoidance and endurance (Lanstyák 2018). In the paper, all of these strategies are subsumed under the 'management' of language problems.

1.2. The authority and role of orthographic codification in Hungary

As will become apparent later, the vast majority of language problems with which the LCS of the HUN-REN Hungarian Research Centre for Linguistics is contacted are related to Hungarian spelling. To understand this fact, it is necessary to have an overview of the situation and role of spelling codification in Hungary.

Spelling is one of the most important issues in Hungarian organised language management, as the level of compliance with spelling codification strongly determines the perceived adequacy of written texts. Various codification styles (Garvin 1993) and institutional frameworks for spelling regulation exist: in addition to the legislative model and

and prefers it to its more recent synonym, *esszenciális olaj* 'essential oil'. The concept of *neutral evaluation* is not explained in the LMT literature. In the present paper, it is applied to situations when a linguistic phenomenon is for some reason salient to the speaker, so the LMT process reaches the first stage (noting), but it is not evaluated either negatively or positively, so at this stage the LMT cycle terminates. For example, it happens when someone makes a mistake in a chat message (e.g. a typo) and realises it, but does not spend time correcting it because the other party can understand what he or she is saying.

publisher-defined orthographies, the academic model is also common (see Baddeley & Voeste eds. 2012). Hungarian orthography has held academic status since 1832, when the first normative spelling rule book was published. Today, the regulation of Hungarian orthography falls under the authority of the Interdepartmental Permanent Committee on the Hungarian Language at the Hungarian Academy of Sciences.

Academic spelling rules are widely accepted and hold significant societal importance. They are not mandatory, but they carry high prestige. Teaching Hungarian orthography rules is especially crucial in teaching Hungarian as a first language. This importance is highlighted by the annual orthographic competitions held at primary, secondary, and university levels.

Although Hungarian orthography is not legally enforced, a significant part of the Hungarian speech community regards academic spelling as an authoritative and unquestionable standard.[2]

2. Language consulting activity in the HUN-REN Hungarian Research Centre for Linguistics

The HUN-REN Hungarian Research Centre for Linguistics has operated a Language Consulting Service since the foundation of the predecessor institute (Research Institute for Linguistics of the Hungarian Academy of Sciences, 1949). The LCS can be contacted by email and telephone with any language problems or questions. Telephone language consulting is available twice a week for four hours. Language consulting by email is available to the general public five working days a week. The latter is obviously the main medium for language consulting, and therefore only email language consulting is discussed in the paper.

2.1. Tools for language consulting

Reference works play a crucial role in the work of language consulting services (Riegel 2007). The HUN-REN LCS also has basic tools at its disposal, including spelling rulebooks and dictionaries,

[2] Beneš et al. (2018: 136–137) report similar attitudes towards spelling codification within the Czech speech community.

monolingual (explanatory) dictionaries (Ittzés et al. 2006–2021; Lipp & Simon 2021) and other dictionaries, descriptive grammars and language usage handbooks. In many cases, real language data is needed to answer language questions. These data can be obtained from corpora. Corpus data can be extracted by direct automatic text processing, or in a more user-friendly way by using corpus query interfaces (Sass 2022). The latter are interfaces that allow searching for words or word combinations with specific properties in large text collections. Thus, the use of corpus query tools is crucial not only in linguistic research but also in applied linguistic practice such as language consulting (Van de Velde & Zenner 2010). The following section briefly describes the Hungarian corpora used by LCS staff in their language consulting activities. It also briefly introduces another tool, the Hungarian Spelling Advisory Portal (HSAP), which is basically a website for managing users' spelling problems.

2.1.1. Hungarian Gigaword Corpus

The Hungarian Gigaword Corpus (henceforth HGC) is a representative corpus of present-day Hungarian, which facilitates the study of Hungarian language varieties in Hungary and beyond its borders. Currently containing 1.5 billion word tokens, the database mostly comprises texts produced in Hungary but it also contains a significant amount of Hungarian language data from beyond the border (Slovakia, Transcarpathia, Transylvania and Vojvodina). The HGC is an annotated corpus offering lemmatisation, part of speech (POS) tagging and morphological analysis for each word (Oravecz, Váradi & Sass 2014).

2.1.2. Hungarian Historical Corpus

The Hungarian Historical Corpus (henceforth HHC) is a collection of texts written between 1772 and 2010 in different genres, containing around 30 million word tokens. During the compilation of the HHC, text samples were selected by professionals (literary historians, historians, mathematicians, etc.) from printed works. In contrast with so-called annotated corpora – in which each word may be accompanied by various types of additional information (e.g. lemma or morphological annotation) –, in the HHC there is no additional information, i.e. the text material can be considered simply as a series of word forms (Sass 2017).

2.1.3. Historical Corpus of Personal Texts (Old and Middle Hungarian corpus of informal language use)

This corpus aims to represent Old and Middle Hungarian (16–18th c.). The text of the corpus has been compiled to approximate as closely as possible the vernacular of the historical periods. The corpus consists of testimonies from court cases and samples from private correspondence. The texts are not only annotated morphologically, but each file also contains additional metadata (Dömötör et al. 2018).

2.1.4. Hungarian Spelling Advisory Portal (HSAP)

Unlike the above corpora, the tool presented in this section is aimed at inquirers rather than linguists and language professionals. However, as it is closely connected to the work of the LCS, a brief description is necessary. The online Hungarian Spelling Advisory portal *Helyesiras.mta.hu* (henceforth HSAP), launched in 2013, offers a range of language technology tools and resources (Váradi 2009; Váradi, Ludányi & Kovács 2014) to alleviate the work of language consultants engaged in solving spelling problems that can be answered by using automated methods. In some cases, use of the HSAP can replace the use of a dictionary or the academic spelling rulebook. Some of the rules of Hungarian normative spelling can be adequately described in a formal language, and turned into algorithms or managed by word lists. Accordingly, the HSAP can provide assistance in the following areas of orthography: writing compound words as one or more words, word-level spell-checking (checking the correctness of certain simple [not compound] words), the spelling of proper names (mainly geographical names), hyphenation, the spelling of numbers and dates, alphabetisation.

2.2. Meta-problems (problem types) in LCS practice

A study of emails received by the LCS between 2012 and 2022 reveals that 82% of the language problems are related to orthography, which clearly shows that normative spelling is highly prestigious in Hungarian society (Ludányi & Domonkosi 2023: 78). A classification of language problems is presented below: meta-problems related to spelling as well as non-spelling issues.

2.2.1. Meta-problems related to spelling

Spelling problems that are reported to the LCS are typically not related to simple, grapheme-level rules but to other, higher-order normative rules. In many cases the inquirers ask for help because a previous adjustment design was unsuccessful as they could not find a relevant rule in the Hungarian spelling rulebook. In this chapter, the most prominent spelling topics in LCS practice are described.

Writing word sequences as one compound word or as separate words is one of the most difficult areas of Hungarian spelling, as shown by the large number of questions on this topic received by the LCS. This is because the relevant spelling rules are determined by the rules of Hungarian grammar. Hence, in order to be able to apply the rules of word breaks, one must have a thorough knowledge of grammar, and recognise the difference between phrases and compounds. Compound words are usually spelt as one word (without spaces) and phrases are normally written as more than one word (with spaces), but this is not always the case. One of the main principles of Hungarian orthography is that compounds must be written without spaces when there is a change of meaning that cannot be deduced from the elements themselves.[3] But it is not always obvious whether there is a change of meaning.

There are special spelling rules for writing multiple compounds (composed of more than two words), which people often have difficulty with, according to inquiries to the LCS. For example, a special rule applies to three-word compounds of more than six syllables.[4] Everyday speakers find these special rules difficult to understand. The LCS helps inquirers develop their adjustment design by explaining the complex rules in a clear and understandable manner. Hyphenation is an alternative to writing as one word and is used, for example, for certain

[3] For example: *kerekes szék* ‘office chair, literally: chair on wheels’, *kerekesszék* ‘wheelchair’.

[4] If a compound consists of more than two words, the syllables must be counted and written in one word up to six syllables, but above six syllables (excluding inflectional suffixes as well as the derivational suffix *-i*) a hyphen must be placed at the main constituent boundary within the compound. However, the “syllable counting rule” causes many difficulties in practice, as speakers are often not familiar with the rule and take the syllable count into account even for two-word compounds, and in other cases it is difficult to decide whether a compound is a two-word or a multi-word compound. For example, it may not be clear whether foreign prefixes (*anti-, kilo-, milli-*, etc.) count as a single word in a compound, cf. *antibiotikum* + *kezelés* ‘treatment with antibiotics’).

“long” compounds (Domonkosi & Ludányi forthc.) or when a compound word contains a proper name.

A common meta-problem is the spelling of new foreign loanwords (e.g. *kapucsínó ~ cappuccino, podcast ~ podkaszt*). For example, about it is often controversial whether they should be written according to the original language’s spelling conventions or according to the Hungarian pronunciation, for which there is no uniform rule. Another common issue is how foreign loanwords should be accompanied by suffixes. Additionally, a frequent meta-problem arises when such loanwords form the first element of a compound, leading to uncertainty about whether the second part should be connected with or without a hyphen.

There are also inquiries regarding the guidelines for transcribing foreign words and expressions into Hungarian. A particular challenge is posed by the transcription of proper names from foreign languages which are not written in the Latin script.

Questions also often arise about the spelling of specific Hungarian proper names. In particular, the inquirers contact the LCS about geographical names. Usually, the problem is how to connect suffixes to proper nouns and how to use them to form compound words.

Inquirers often have problems with the use of commas, especially before the conjunctions *és* ‘and’, *vagy* ‘or’ and *mint* ‘as, than’. There are also questions about word hyphenation rules, spelling of numbers and dates.

While the LCS primarily manage general language problems, it is also frequently approached with questions on specialised languages. Certain professions have their own specific spelling dictionaries; this applies to fields such as economics, medicine, technology, chemistry, military affairs, zoology, and botany. Providing answers to spelling questions related to technical terms can be particularly challenging because applying the spelling rules of a given specialised language requires a thorough knowledge of the subject matter as well as the practice of the specialised language (which is why the LCS cannot always help with these problems).

The LCS often receives feedback on the HSAP, with inquirers reporting on their experience with the portal while attempting to develop an adjustment design. In particular, when the portal does not provide an (unambiguous) answer and/or the inquirers do not agree with the website’s answer, they contact the LCS, thus starting another language management cycle.

2.2.2. Meta-problems not related to spelling

Two further types of inquiries, not related to normative spelling, can be distinguished: 1) emails asking for advice on language use, 2) requests for language information rather than for language problem management.

The main types of language problems related to language use concern the following areas: the naming of new phenomena, new terms; finding the Hungarian equivalent of foreign words and expressions; word variants; the existence of words; finding the meaning of words; questions on pronunciation; questions related to linguistic politeness, in particular in the area of address forms (e.g. how to address unknown women in email); suggestions for "introducing" new words; conscious creation of new words; language superstitions (Ludányi & Domonkosi 2023: 82).

Common meta-problems of emails requesting language information include etymological issues (for example, the origin of certain words or phrases) and the first occurrences of certain words.

3. Description of the research

3.1. Research question and hypotheses

The study is a self-reflective study of how staff members of the Language Consulting Service (LCS) of the HUN-REN Hungarian Research Centre for Linguistics have used corpus data to answer questions received by the LCS over the past ten years. The research was motivated by the idea that language consulting could benefit from a study of past corpus use patterns. The following research question was formulated:

(RQ) What are the language problems noted by inquirers that LCS staff used corpus data to manage when developing an adjustment design?

To answer the research question, the following two hypotheses were formulated.

(H1) LCS staff rely on real language data more often when answering language use questions than when answering spelling questions.

(H2) When answering spelling questions, language consultants examine spelling practice using corpus data 1) in cases of inconsistency, 2) and in cases where spelling codification is not available.

3.2. Material and method of the analysis

The analysis presented in this study is based on LCS data, namely questions received by email and the corresponding answers given by language consultants. An anonymised database of over 10,000 emails is currently under development (Domonkosi & Ludányi 2023).[5] The database contains emails received since September 2011.

Each email sent to the LCS receives codes from a structured system of labels in parallel with the production of a response. The current primary labelling system basically classifies problems according to the type of linguistic phenomenon (e.g. spelling problems and subproblems within them, language use issues such as word formation, linguistic politeness, stylistic, etymological issues). The annotation of emails is carried out partly manually and partly automatically. Automatic annotation can be set up by using the mail system's filter conditions, so that when an email matches a filter condition (e.g. it contains certain keywords), the relevant label is automatically added.

At the top of the label hierarchy are two labels: 1) spelling questions, 2) non-spelling questions. In addition to the primary annotation identifying the type of language problem, LCS staff are also working on the development of additional annotation methods during the subsequent manual coding and processing of emails (Domonkosi & Ludányi 2023). The emails are also annotated from the point of view of whether the language consultant used corpus data to answer them. For the present analysis, emails with the annotations "corpus use/HGC" and "corpus use/HHC" were selected.[6]

The frequency of use of the two corpus query interfaces is clearly shown by the ratios of the emails selected for the analysis: in answering a total of 118 emails out of the material containing more than 10,000 emails, the LCS staff relied on the representative corpus of the Hungarian language, the HGC, and much less often, in only 13 cases, to the

[5] Data of those turning to the LCS of the HUN-REN Hungarian Research Centre for Linguistics are handled in accordance with regulations. The material of emails is included in the database in an anonymised form, and inquirers are informed about the use of the linguistic material of their questions for purposes of academic research.

[6] For ease of understanding, the Hungarian codes (labels) have been translated into English. The Hungarian equivalents of the labels "corpus use/HGC" and "corpus use/HHC" are „korpuszhasználat/MNSz" and „korpuszhasználat/MTSz", respectively.

HHC. The figures also include overlaps, as in seven cases the LCS staff member, applying the "principle of cooperation" (Sass 2022: 606–608), used both corpora to answer the question.

It is important to note that there is a growing trend in the use of corpus query interfaces in LCS practice. While in the first half of the 2010s, corpus use was only sporadic, from 2015 onwards it has gradually become more frequent,[7] as shown in Figure 1. (The last email of the research material is from May 2023, which explains the low number of data for 2023.)

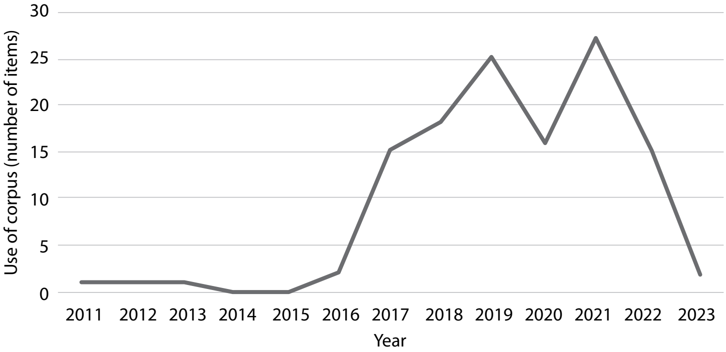

Figure 1. Distribution of LCS responses based on corpus data by year (November 2011 – May 2023).

The analysis material was compiled from a total of 124 question-answer pairs obtained by preliminary filtering. This material has been further annotated. The correspondence was inserted into a MS Excel spreadsheet. After the removal of personal data of the inquirer, each correspondence was assigned a unique identifier. The resulting corpus includes:

- the date of receipt of the inquiry
- a concise description of the language problem (e.g. whether *hotele* or *hotelje* 'his/her hotel' was the correct/standard third person singular possessive form of *hotel* 'hotel')
- initials of the responding LCS staff member
- the full text of the question email and the reply email

[7] This is clearly related to an increase in the number of research group/LCS staff. Since 2015, an increasing number of young researchers have joined the group.

– the corpus used (HGC/HHC or Historical Corpus of Personal Texts)
– and any comments.

The table also contains an attribute named "Spelling / Non-spelling", which can take only one of the values "Spelling" and "Non-spelling".

The generalisation of problem tokens has been completed and the main meta-problem categories have been created. When at least two similar problem instances occurred, they were classified into one type. For example, when the language problem was caused by the difficulty in forming the plural form of the noun *mokaszin* 'moccasin' (#hgc60)[8] and the instrumental case form of the noun *nokedli* 'dumpling' (#hgc80), the problem was generalised and a separate code (meta-problem category) was created for the difficulty inherent in suffixing so-called disharmonic stems ending in a front unrounded vowel (*e*, *é*, *i*, *í*). Because of the diversity of problem instances, especially in the category "Non-spelling", it was necessary to create an "Other" type for cases where the generalisation of problem instances into meta-problems was not feasible.

3.3. Results of the analysis

The representative corpus of the Hungarian language, the HGC, containing 1.5 billion tokens, is the most commonly used corpus in the work of the LCS. Data from the HHC (which contains texts from 1772–2010) are much less frequently relied upon by LCS staff in formulating their response. Of the 13 questions in this group, HHC was used in 7 cases by the LCS in addition to the HGC, and in one case the Historical Corpus of Personal Texts (Dömötör et al. 2018) was used in addition to the HHC. Table 1 shows the overall distribution of the results obtained.

[8] In each case, the sequence of characters starting with a hashtag in parentheses represents the unique identifier of the language problem. For ease of understanding, the identifiers have been changed to the English abbreviation of the corpus name: for example, #hgc83 was originally #mnsz83.

Table 1. Distribution of language consultant responses based on corpus data.

	Spelling	Non-spelling	Total
HGC (including use in combination with other corpora)	26	92	118
HHC only	2	3	5
HHC and Historical Corpus of Personal Texts	0	1	1
Total	28	96	124
%	~8000	~2000	~10000
	~0,35%	4,8%	~1,2%

3.3.1. Using corpora to manage non-spelling problems

Out of 118 email responses based on HGC data, language consultants used the corpus data in 92 cases (78%) for non-spelling (typically language use) issues. Figure 2 presents frequency data for meta-problems concerning language use.

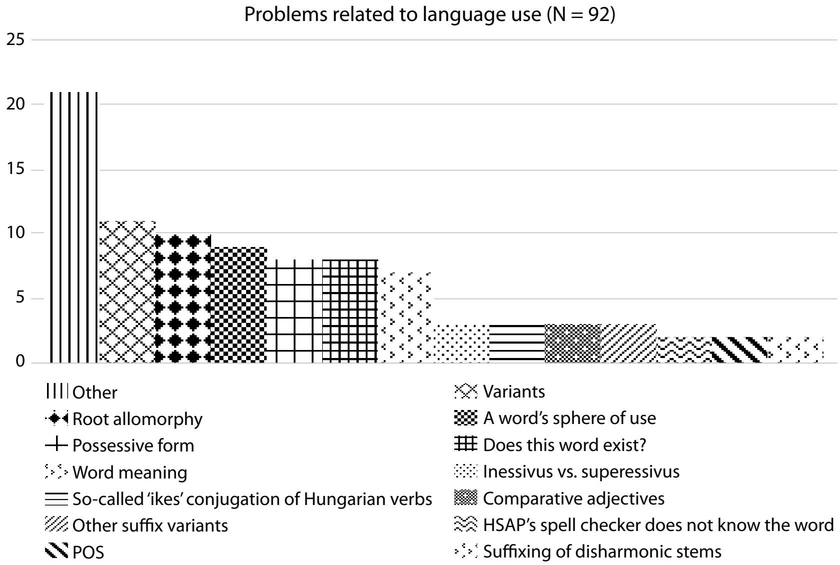

Figure 2. Distribution of meta-problems related to language use, answered by HGC data. The category Other includes only instances of problems occurring once.

The data from the analysis show that LCS staff rely on corpus data most often to answer questions about lexeme variations (11 questions). In the annotation process, based on Lőrincz, Lőrincz & Török (2021: 292), those language units were considered lexical variants “that are similar in form (they show only partial, phonetic differences in form that do not constitute a morpheme-level deviation), and have completely identical referential (denotative) lexical and grammatical meanings, but their pragmatic meanings are different”. Such variants are, for example *irattárazás* ~ *irattározás* ‘archiving’ (#hgc12), *felülvizsgálat* ~ *fölülvizsgálat* ‘review’ (#hgc15).

Variability is also present at the morpheme level and the data suggest that it is the topic of a common question type. Questions about the variations of lemmas/stems can be considered a meta-problem exemplified by e.g. *olajos magok* ~ *magvak* ‘oilseeds’ (#hgc23) and *számlát igényelek* ~ *igénylek* ‘I request an invoice’ (#hgc98) (10 questions). And at the level of affix morphemes, difficulties arise in particular with the formation of the possessive form of nouns (8 questions): *hangukat* ~ *hangjukat* ‘voice.px(3pl).acc’, (in accusative)’ (#hgc96), *hotele* ~ *hotelje* ‘hotel.px(3sg)’ (#hgc116). Other, less common but persistent meta-problems of morpheme-level variability include the so-called ‘ikes’ conjugation (Kozmács 2020) (*nem kések el* ~ *nem késem el*[9] ‘I am not late’, #hgc97), the use of inessive vs. superessive suffixes, especially for settlement names (*Balmazújvárosban* ~ *Balmazújvároson*, #hgc29), and the difficulties of choosing between different suffix variants (*orrszarvúk* ~ *orrszarvúak* ‘rhinoceroses’, #hgc106).

Phonemic variability causes language problems most often in cases where some words (so-called disharmonic roots) ending in a front unrounded vowel (*e*, *é*, *i*, *í*) need to be affixed. On the basis of their phonological behaviour, in some cases, these vowels may be considered neutral. They behave transparently in terms of backness harmony, combining in large number to form mixed roots. Many of the stems with back and neutral vowel structures are grammatical with both front and back suffixes (vacillating stems) (Ringen & Kontra 1989; Benkő 2014):

[9] In the paradigm of verbs whose third person singular form ends in *-ik,* the more traditional, prestige-valued first-person singular form has an *-m* personal suffix (*m*-variant: *késem* ‘I am late’). The variant ending in *-k* is a more recent development (*kések* ‘I am late’) motivated by the analogy of verbs with *-ik*-less conjugation.

mokaszineket ~ *mokaszinokat* ‘moccasin.PL.ACC’ (#hgc60), *nokedlival* ~ *nokedlivel* ‘dumpling.INSTR’ (#hgc80).

The other group of meta-problems concerning language use, answered by the HGC data, is not related to linguistic variability. One of the most frequent meta-problems is when inquirers ask for information about the value of use (9 questions), meaning (7 questions) or existence of certain lexemes (rarely phrases) (8 questions).

For example, an employee of an insecticide company wanted to know if the term *büdösbogár* (literally ‘stink bug’), which is used in their everyday professional language, could be used to communicate with customers about the *Halyomorpha halys* ‘brown marmorated stink bug’ (#hgc99, value of use of a word as a meta-problem). Another inquirer asked whether the adjective *szaftos* ‘juicy’ could be used to describe a succulent cake, as *szaft* is typically used to mean ‘fatty, spicy meat juice’ (#hgc113, word meaning as a meta-problem).

LCS staff also often rely on HGC data to answer questions such as “does this word exist?”. In this category, inquirers tend to raise questions either about rare and/or novel words used by small communities of practice (e.g. *blőr* ‘blurring part of a photograph’, #hgc87) or about words underlined by spell checker tools and marked as unknown (*unkonvencionális* ‘unconventional’, #hgc47). A related case was classified as a specific meta-problem, when the inquirer did not ask for information about the existence of a word, but informed the LCS that the spell checker tool of the HSAP marked certain words as unknown. In these cases, the purpose of using the corpus in the consulting strategy is to confirm the existence of the word, to present frequency data and to rely on them to determine the value of use of the word.

Another morpheme-level meta-problem, not related to variability, concerns the comparative degree of intensifying adjectives (3 questions). For example, an inquirer asked whether the adjective *létfontosságú* ‘vital, essential’ had a comparative form (#hgc78, #hgc108). A rare meta-problem (2 questions) is when inquirers ask for information about the part of speech of certain lexemes (e.g. a mother asked about the part of speech of *haladó* ‘advanced’ for helping with the grammar homework of her child, #hgc63).

Problems that cannot be abstracted were placed in the “Other” category, and these problems show a large variety.

Compared to the Hungarian Gigaword Corpus, the Hungarian Historical Corpus has been used significantly less often in the development of the adjustment design. In terms of the ratio of spelling to non-spelling questions, HHC was used by LCS staff in 11 non-spelling cases.

In the analysed material, there are 5 language problems in the management of which the consultants only used data collected from the HHC. These are individual, specific problems that cannot be merged into meta-problems: for example, the first occurrence of the word *sziesztа* ‘siesta’ (#hhc4) or the spelling of the surname of a famous Hungarian actress: *Laborfalvy* ~ *Laborfalvi Róza* (#hhc5).

3.3.2. Using corpora to manage spelling problems

In addition to language use problems, LCS staff also use HGC data less frequently, only in 26 cases (22%) when answering spelling questions language consultants used the corpus data. Figure 3 shows the distribution of spelling meta-problems.

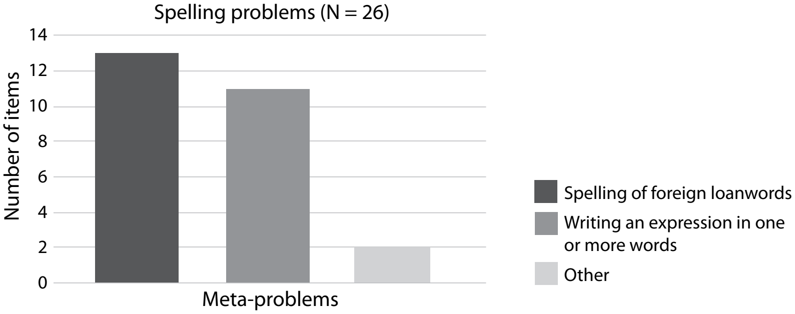

Figure 3. Distribution of meta-problems related to spelling, as answered by HGC data.

Compared to language use problems, spelling meta-problems do not show much variation. The two main meta-problems that LCS staff used corpus data (always HGC data) to manage concerned the spelling of foreign loanwords (13 questions) and writing in one or two words (11 questions). The Hungarian Historical Corpus was used by LCS staff in only 2 spelling cases, and only for control purposes, including combined use with other corpora.

One of the most common ways to increase vocabulary is to borrow words from foreign languages. The spelling of new foreign (especially English) loanwords typically causes a spelling problem where it seems essential to investigate spelling practice. In all cases, the problem is caused by the choice between spelling that follows the source language and spelling that follows the Hungarian pronunciation.

In Hungarian orthography, the spelling of loanwords can be of three types: 1) only spelled according to the source language in the dictionary (e.g. *déjà vu*, *couchette*, *flow*), 2) only spelled according to Hungarian pronunciation in the dictionary (e.g. *randevú* [< *rendez-vous* 'an arranged date'], *lájkol* 'to like', *szelfi* 'selfie', 3) alternate spelling, spelling variations in the dictionary, i.e. both spellings are codified (*e-mail* ~ *ímél*, *chat* ~ *cset*). There are examples for all three types in spelling dictionaries. What causes a language problem for language users is when the newly borrowed neologisms do not yet have a codified spelling, because it is difficult to decide when to spell the loanword following the source language, according to the Hungarian pronunciation, or with alternate spellings. There are several such questions in the material analysed: *avatar* ~ *avatár* (#hgc28), *gamer* ~ *gémer* 'a person who plays computer games more seriously than as a hobby' (#hgc45), *verniszázs* ~ *vernisszázs* 'opening ceremony of a painting exhibition' (#hgc89), *tweetel* ~ *twittel* 'post a tweet on the Twitter social media site'(#hgc100).

As mentioned before, writing compound as single word or a sequence of separate words is one of the most difficult areas of Hungarian spelling. This is evidenced by the large number of such inquiries received by the LCS. To answer some of these questions, LCS staff also consult corpus data. For example, should the idiomatic expression *testreszab* ~ *testre szab* 'customise software (or anything else) to the user's needs [originally, tailor something to someone's body]' (#hgc92) be written in one word or two words?

4. Case study: the verb *tervez* 'plan' in the auxiliary verb structure

A teacher asked the LCS to correct the language use of her former pupil who, after inviting her former teacher to an event, asked: *[Keresztnév] bácsi* ***el tervez jönni****?* 'Uncle [first name], **are you going to come**?' This pattern involves the verb *tervez* 'plan' in what can be called the

Hungarian auxiliary construction, which (in default word order) has an auxiliary(-like) verb inserted between a so-called preverb and the infinitive to which the latter lexically belongs (here, *el* is the preverb, also referred to as a verb modifier, which lexically belongs to the verb *jön* 'come', with *eljön* meaning something like 'come along').

In his email, the inquirer expressed a negative attitude towards this use of the verb *tervez*, considering it as incorrect. However, he also acknowledged that for his student this type of grammatical construction was probably as natural as using the future auxiliary *fog* 'will, be going to' with a similar structure: *el fogsz jönni?* 'Are you going to come?'

In the LCS response, it was confirmed that the construction *el tervez jönni* 'is going to come' is indeed similar to the pattern with the future auxiliary verb *fog* 'will'. As a first step in the consulting strategy, descriptive grammars were reviewed. The verb *tervez* also follows the verb modifier without stress in a neutral positive declarative sentence (*én el tervezek jönni* 'I am going to come'), and its function is to indicate a distance from reality (Imrényi 2013: 126), i.e. the process of coming along is portrayed as a future possibility depending on someone's intentions rather than a matter of certainty. Therefore, *el tervez jönni* 'he is going to come' can be clearly identified as an instance of the Hungarian auxiliary construction. The use of *tervez* as an auxiliary(-like) verb has also been confirmed by a recent corpus-driven study (Kalivoda & Prószéky 2024: 10).

The second step in the consulting strategy is to search for corpus data similar to the structure *én el tervezek jönni* 'I am going to come'. What is needed here is real language data where *tervez* is embedded between the preverb and the infinitive. In order to obtain suitable matches, the first step is to determine the most specific surface element of the linguistic phenomenon to be investigated (Sass 2022: 603), which in this case is the verb *tervez* itself. In HGC, a 3-step search was performed to obtain the desired linguistic data.

After searching for the verb *tervez* as a lemma, the resulting concordance was filtered in two steps to obtain the corresponding matches (Sass 2022: 609–610). A query was formulated to have a preverb to the left of the key word in context (search window set to −1, −1). For the second filtering, the search window was set to 1,1 and the POS was set to infinitive, so that the infinitive is to the right of the hit, exactly 1 word away.

In order to prove the hypothesis that this is a recent linguistic phenomenon, a search of the HCC containing older texts was also performed. As HCC is not annotated morphologically or otherwise, relevant linguistic data has to be extracted in a different way (Sass 2017). To collect linguistic data, a more complex query formulation was needed, using regular expressions (Mertz 2003) in Corpus Query Language (CQL) (Sketch Engine Team 2015).

The first step was to search for the verb *tervez*, but due to the lack of morphological annotation, the form (regular expression) "tervez.*" was used to retrieve all conjugated forms. Then, the search window was narrowed to –1, –1 to include the most frequent preverbs to the left of the hit, 1 word distance apart: "(el|ki|be|fel|le|át|meg)". The five matches obtained were then filtered again to find an infinitive to the right of the hit, 1 word away, with the main feature being the *-ni* ending: ".*ni". As the search did not find any hits, i.e. the verb *tervez* is not found in the auxiliary verb structure in texts from 1772–2010, it seems likely that this is indeed a recent linguistic phenomenon.

As a next step in the consulting strategy, conclusions were drawn from the available corpus data obtained from the HGC. After further refining the search, manually filtering out false matches, and searching by sub-corpora, it was found that the use of the verb *tervez* in the auxiliary construction is most common in press language (23 matches out of 43 in total), also present in scientific language use, albeit in smaller numbers (6 matches) and in personal posts on social networking sites (5 matches), and rare or absent in other registers.

Overall, therefore, the corpus data suggest that the teacher contacting the LCS was right in observing that the use of *tervez* as an auxiliary verb is not part of the more formal, more polished registers, as it appears more often in the more casual style of press language utterances. However, as it is also used in texts closer to the scientific style, it is conceivable that its use in more formal text types will become more common.

5. Discussion

The research sought to find out what types of language problem inquiries prompted LCS staff to use corpus data when developing an adjustment design. The results suggest that both hypotheses have been confirmed.

(H1) A systematic analysis of the correspondence of the Hungarian Research Centre's LCS shows that performing corpus queries is more prominent as a strategy for answering questions about language use. In terms of sheer numbers, the majority of LCS inquiries are related to the rules of Hungarian spelling (80–90%), and the number of language use questions is much smaller in comparison (10–20%). Proportionally, though, the use of corpus query interfaces is less common in answering the large number of spelling questions (cf. Table 1). The reason for this can be found in the fact that the purpose of managing the two types of problems, spelling and language use, is substantially different. In the case of spelling problems, the intention of the inquirers is to conform to the academic spelling norm,[10] and thus the basic aim of the LCS staff is to communicate that norm. Hence, in most cases, there is no need to examine existing spelling practice on the basis of corpus data. Occasionally, especially in the area of writing compound words as single word or a sequence of separate words, the academic spelling norm is obsolete as it only defines the default case, which requires the use of a space. A good example for this is *kerekesszék* 'wheelchair' in footnote 3, which was previously only spelled with a space in the spelling dictionary (*kerekes szék*), even though corpus data clearly demonstrated that in the usage *kerekesszék* 'wheelchair' it is consistently spelled without a space. Therefore, when the new spelling rules and dictionary were published, both *kerekes szék* 'office chair, literally: chair on wheels' and *kerekesszék* 'wheelchair' were added to the spelling dictionary. Such a change of meaning can also be seen in the expression *testre szab* 'tailor something to someone's body' ~ *testreszab* 'customise software (or anything else) to the user's needs', but now only the variant without space is included in the spelling dictionary.

Although the language ideology of LCS inquirers is strongly influenced by the "correct/incorrect" classification of linguistic expressions (Woldt 2010), the aim of LCS staff in answering such questions goes beyond communication of the norm. In all cases where an inquirer asks about the general correctness of a linguistic form, the LCS response

[10] In some cases, this intention is explicitly communicated to the LCS by the inquirers themselves. Otherwise, it follows directly from the situation of Hungarian spelling regulation and the high social prestige of spelling codification (see chapter 1.2). When the inquirers' intention is not to conform to academic regulations (which rarely happens) but rather to criticise them, they usually express this clearly in their letter.

reflects the fact that the situation is more complex than a question of “correctness” or “incorrectness”. The primary aim of language use problem management in LCS practice is to present the use value of the linguistic phenomenon in question based on linguistic data, research and corpora. Language use issues are addressed by building on the relevant linguistic data, making the data transparent, taking into account register- and context-dependent variation and the perspectives of language users. The consultants are conscious of the language ideologies behind their work (Domonkosi & Ludányi forthc.; Ludányi & Domonkosi 2023).

However, out of the total number of language use questions in the LCS database (~2000 emails), corpora were consulted in only 124 cases. This shows that relying on corpus data does not seem to be absolutely necessary for the management of language use problems. Indeed, there are language use problems which can be successfully managed with the help of dictionary data, relevant chapters of descriptive grammars and language use manuals, and/or available research results. For example, when answering questions on linguistic politeness, corpus use is not part of the consulting strategy (for a discussion of possible reasons, see Domonkosi & Ludányi 2023). Crucially though, refraining from the use of corpora in such cases is not the result of a traditional understanding of linguists’ role or the adoption of a prescriptive approach. Rather, sometimes it is simply sufficient to look up the (new) meaning of a lexeme in the Hungarian monolingual dictionary (Ittzés et al. 2006–2021), which is corpus-based and constantly updated online.

The results of the analysis also show that the language use problems that the LCS staff used corpus searches to address can be divided into two groups. One group of problems is related to language variability, with the most frequent difficulties arising in the use of lexeme- and morpheme-level variants. The other group of language use questions answered on the basis of corpus data concerns the value of use, meaning, or existence of certain lexemes or phrases.

(H2) The analysis has also shown that the use of a corpus is part of the management strategy for normative spelling problems only when spelling codification is absent or inconsistent. The data suggest that conducting corpus searches was necessary in a subset of cases where the spelling problem was caused by a discrepancy between codification and customary usage. These problems occurred in the area of writing compound words as single word or a sequence of separate words. The codified spelling norm was generally two-word spelling, but the usage preferred one-word spelling (e. g. *testre szab* ~ *testreszab*, ‘customize’, *jó gyakorlat* ~ *jógyakorlat* ‘good practice’.

Based on the data analysed, the use of a corpus is part of the strategy for language consulting in the absence of spelling codification. In particular, for spelling issues of new loanwords, it is typical for consultants to examine spelling practice by using corpus data. This is due to the fact that there is no general rule in Hungarian orthography for the spelling of foreign loanwords without dictionary codification.

The Hungarian spelling rulebook gives no general rule as to when to write these neologisms following the source language, when to write them following the Hungarian pronunciation, and when to allow for alternate spelling. In such cases, each case has to be examined individually, based on a number of criteria. Frequency, prevalence, common spelling practice and “how much the term is perceived as Hungarian” by the language users are important criteria. In such cases, the use of corpora is an indispensable part of the consulting strategy.

6. Summary and outlook

The paper analysed answers to inquiries received by the LCS with the aim of examining which meta-problems invite the use of corpus query interfaces. The results show that the use of corpus query interfaces is an effective method for managing complex language problems and for describing the usage of particular expressions. On the other hand, corpus data are rarely needed for answering questions on orthography, which constitute the bulk of inquiries that the LCS receives. In answering spelling questions, the basic aim is to communicate the spelling norm, thus the examination of usage is only necessary when spelling codification is absent or inconsistent. The paper presented a case study on a specific language use issue from the database, highlighting the strategy used in the development of language advice in detail, with a particular focus on the appropriate use of corpus query interfaces.

However, the proportions of question types show that overall, the use of corpus data in the formulation of a response was relatively uncommon (124 cases), i.e. ~4.8% of the language use questions, and

~1.2% of the total. The data suggest that in ~95% of the language use questions, corpus use was not part of the language consulting strategy. Hence, there is an important future task to systematically investigate which tools and reference works are generally used by LCS staff to answer various types of language use questions.

As an outlook, the results of the present research can be seen as a contribution to the creation of a state-of-the-art writing assistant aiding language users that will be developed in the future, exploiting advances in language technology. Such a tool can be visualised as a digital handbook in which corpora and databases are integrated. Authentic language data is indeed essential for language consulting. Therefore, the Hungarian Usage and Language Consulting Research Group aims to create a language problem database in the future, partly on the basis of LCS inquiries.

References

Baddeley, Susan & Anja Voeste (eds.). 2012. *Orthographies in early modern Europe*. Berlin/Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110288179.1.

Beneš, Martin, Martin Prošek, Kamila Smejkalová & Veronika Štěpanová. 2018. Interaction between language users and a language consulting center: Challenges for language management theory research. In Lisa Fairbrother, Jiří Nekvapil, & Marián Sloboda (eds), *The Language Management Approach: A Focus on Research Methodology*, 119–140. Berlin: Peter Lang.

Benkő, Ágnes. 2014. Vacillating stems in Hungarian. *The Even Yearbook* 11. http://seas.elte.hu/w/!even/_media/14be.pdf (9 April, 2024).

Biere, Bernd Ulrich & Rudolf Hoberg (eds.) 1995. *Bewertungskriterien in der Sprachberatung.* Tübingen: Narr.

Domonkosi, Ágnes & Zsófia Ludányi. 2023. Politeness metadiscourses in the practice of language consulting. *Studia Linguistica Hungarica* 35. 24–37. https://doi.org/10.54888/slh.2023.35.24.37.

Domonkosi, Ágnes & Zsófia Ludányi. (forthc.) Language consulting in Hungary: a case study on the practices of the Hungarian Language Consulting Service.

Dömötör, Adrienne, Katalin Gugán, Attila Novák & Mónika Varga. 2018. Creation of an annotated corpus of Old and Middle Hungarian court records and private correspondence. *Language Resources and Evaluation* 52. 1–28. https://doi.org/10.1007/s10579-017-9393-8.

Garvin, Paul L. 1993. Styles of codification. *Brno Studies in English* 20. 17–21.

HGC = Hungarian Gigaword Corpus. http://corpus.nytud.hu/mnsz (9 April, 2024).

HHC = Hungarian Historical Corpus. http://clara.nytud.hu/mtsz/ (9 April, 2024).

Historical Corpus of Personal Texts. https://tmk.nytud.hu (9 April, 2024).

HSAP = Hungarian Spelling Advisory Portal. https://helyesiras.mta.hu/ (9 April, 2024).

Ittzés, Nóra et al. 2006–2021. *A magyar nyelv nagyszótára I–VIII.* [Comprehensive Dictionary of Hungarian]. Budapest: Nyelvtudományi Kutatóközpont.

Imrényi, András. 2013. The syntax of Hungarian auxiliaries: a dependency grammar account. In Eva Hajičová, Kim Gerdes & Leo Wanner (eds.), *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*, 118–127. Prague: Charles University.

Jernudd, Björn Holger & Jiří Václav Neustupný. 1987. Language planning: for whom? In Lorne Laforge (ed.), *Actes du Colloque international sur l'aménagement linguistique. Proceedings of the International Colloquium on Language Planning,* 69–84. Québec: Les Presses de L'Université Laval.

Kalivoda, Ágnes & Gábor Prószéky. 2024. Hungarian auxiliaries revisited. *Acta Linguistica Academica* 71(1–2). 202–218. https://doi.org/10.1556/2062.2023.00701.

Kimura, Goro C. & Lisa Fairbrother. 2020. Reconsidering the language management approach in light of the micro-macro continuum. In Goro C. Kimura & Lisa Fairbrother (eds)., *Language Management Approach to Language Problems: Integrating Macro and Micro Dimensions*, 255–267. Amsterdam (Philadelphia): John Benjamins. https://doi.org/10.1075/wlp.7.13kim.

Kopecký, Jakub. 2022. Divergent interests and argumentation in Czech Language Consulting Center interactions. In Marek Nekula, Tamah Sherman & Halina Zawiszová (eds.). *Interests and Power in Language Management*, 73–99. Berlin: Peter Lang.

Kozmács, István. 2020. A receding paradigm as a tool of language discrimination. *Hungarian Studies: A Journal of the International Association for Hungarian Studies and Balassi Institute* 34(1). 108–119. https://doi.org/10.1556/044.2020.00010.

Lanstyák, István. 2014. On the process of language problem management. *Slovo a slovesnost* 75(4). 325–351.

Lanstyák, István. 2018. On the strategies of managing language problems. In Lisa Fairbrother, Jiří Nekvapil & Marián Sloboda (eds.), *The Language Management Approach: A Focus on Research Methodology*, 67–97. Berlin: Peter Lang.

Lengar Verovnik, Tina & Helena Dobrovoljc. 2022. Revision of Slovenian Normative Guide: Scientific Basis and Inclusion of the Public. *Slovene Linguistic Studies* 14. 183–205. https://doi.org/10.3986/sjsls.14.1.07.

Lipp, Veronika & László Simon. 2021. Towards a new monolingual Hungarian explanatory dictionary: overview of the Hungarian explanatory dictionaries. *Studia Lexicographica* 15(29). 83–96.

Lőrincz, Gábor, Julianna Lőrincz & Tamás Török. 2021. Language variativity and contact variants in the Hungarian and international literature Vestnik Ugrovedenia / *Bulletin of Ugric Studies* 11(2). 292–300. https://doi.org/10.30624/2220-4156-2021-11-2-292-300.

Ludányi, Zsófia. 2020. Language consulting: A brief European overview. *Eruditio – Educatio* 15(3). 25–47. http://doi.org/10.36007/eruedu.2020.3.025-047.

Ludányi, Zsófia & Ágnes Domonkosi. 2023. Language consulting and language management from the perspective of the Hungarian language Consulting Service. *Taikomoji kalbotyra* 20. 74–88. https://doi.org/10.15388/Taikalbot.2023.20.6.

Mertz, David. 2003. Regular expressions. Introduction to the Tutorial. In David Mertz: *Text Processing in Python.* 1st edition. Addison-Wesley. https://gnosis.cx/publish/programming/regular_expressions.html (9 April, 2024).

Nekvapil, Jiří. 2006. From Language Planning to Language Management. *Sociolinguistica* 20. 92–104. https://doi.org/10.1515/9783484604841.92.

Nekvapil, Jiří. 2009. The integrative potential of Language Management Theory. In: Jiři Nekvapil & Tamah Sherman (eds.), *Language management in contact situations: Perspectives from three continents*, 1–11. Frankfurt am Main: Peter Lang.

Nekvapil, Jiří. 2012. Some thoughts on "noting" in language management theory and beyond. *Journal of Asian Pacific Communication* 22(2). 160–173. https://doi.org/10.1075/japc.22.2.02nek.

Neustupný, Jiří V. 2003. Japanese students in Prague: Problems of communication and interaction. *International Journal of the Sociology of Language* 162. 125–143. https://doi.org/10.1515/ijsl.2003.033.

Oravecz, Csaba, Tamás Váradi & Bálint Sass. 2014. The Hungarian Gigaword Corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck et al. (eds.), *Proceedings of Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, 1719–1723. Reykjavik: ELRA.

Prošek, Martin. 2020. Processes of language enquiries. The case of the Prague Language Consulting Service. In Goro C. Kimura & Lisa Fairbrother (eds.), *Language Management Approach to Language Problems: Integrating Macro and Micro Dimensions*, 197–213. Amsterdam: John Benjamins. https://doi.org/10.1075/wlp.7.10pro.

Riegel, Mareike. 2007. *Sprachberatung im Kontext von Sprachpflege und im Verhältnis zu Nachslagewerken. Unter besonderer Beachtung der Sprachberatungsstelle des Wissen Media Verlages*. Freiburg i. Br.: Albert-Ludwigs-Universität PhD Dissertation.

Ringen, Catherine O. & Miklós Kontra. 1989. Hungarian neutral vowels. *Lingua* 78. 181–191. https://doi.org/10.1016/0024-3841(89)90052-1.

Sass, Bálint. 2017. A kibővített Magyar Történeti Szövegtár új keresőfelülete [Querying corpora: The new search interface of the expanded Hungarian historical corpus]. In Forgács Tamás, Németh Miklós & Sinkovics Balázs (szerk.), *A nyelvtörténeti kutatások újabb eredményei IX.*, 267–277. Szeged: SZTE Magyar Nyelvészeti Tanszék.

Sass, Bálint. 2022. Principles of corpus querying: A discussion note. *Acta Linguistica Academica* 69(4). 599–614. https://doi.org/10.1556/2062.2022.00581.

Sketch Engine Team. 2015. *CQL – Corpus Query Language*. https://www.sketchengine.eu/documentation/corpus-querying (9 April, 2024).

Velde, Freek van de & Eline Zenner. 2010. "Pimp my Lexis": het nut van corpusonderzoek in normatief taaladvies. In Els Hendrickx, Karl Hendrickx, Willy Martin, Hans Smessaert, William van Belle & Joop van der Horst (eds.), *Liever meer of juist minder? normen en variatie in taal*, 51–68. Gent: Academia Press.

Váradi, Tamás. 2009. Bringing language technology to the masses. Some thoughts on the Hungarian online spelling dictionary project. In Dana Hlaváčková, Aleš Horák, Klára Osolsobě & Pavel Rychlý (eds.), *After half a century of Slavonic natural language processing*, 227–230. Brno: Tribun EU.

Váradi, Tamás, Zsófia Ludányi & Réka Kovács. 2014. Géppel segített helyesírás. A helyesírás.mta.hu portál készítéséről. [Computer aided spelling advice. Designing the helyesiras.mta.hu site]. *Modern Nyelvoktatás* 20(1–2). 43–58. https://helyesiras.mta.hu/ (9 April, 2024).

Vranjek Ošlak, Urška. 2023. Language counselling: Bridging the gap between codification and language use. *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics* 14(1). 149–173. https://doi.org/10.12697/jeful.2023.14.1.05.

Woldt, Claudia. 2010. Metasprache in der tschechischen Sprachberatung – Sprachbewusstsein und seine Reflexe in Laienanfragen und Expertenantworten. *Zeitschrift für Slawistik* 55(2). 176–190. https://doi.org/10.1524/slaw.2010.0014.

Kokkuvõte. Zsófia Ludányi: Korpuste roll HUN-REN Ungari keeleteaduse uurimiskeskuse keelenõustamisteenistuse töös. Keelekorraldusteooriale tuginedes tutvustab artikkel korpusandmete kasutamist keelenõustamisel Ungari Keeleteaduse Uurimiskeskuse HUN-REN keelenõustamisteenistuse (LCS) tegevuse näitel. Keelenõuteenuse arendamisel on oluline arvestada keeleteadusliku taustaga, rakendada andmepõhist lähenemist ning teatud juhtudel võimaldada juurdepääs konkreetsete keeleprobleemide lahendamiseks vajalikele keeleandmetele. Uuring põhineb LCS-i ligikaudu 10 000 e-kirjast koosneval andmebaasil ning analüüsib küsimusi, millele vastamisel keelenõustajad kasutasid korpuspõhiseid meetodeid (124 päringut). Tulemused näitavad, et korpusepäringute kasutamine on tõhus keeruliste keeleküsimuste lahendamisel ja spetsiifiliste väljendite kasutuse kirjeldamisel. Samas on korpusandmete kasutamine harva vajalik ortograafiaga seotud küsimuste puhul, mis moodustavad suure osa LCS-ile esitatavatest päringutest. Õigekirjaküsimustele vastamisel on peamine eesmärk edastada kehtiv õigekirjareegel, seega on kasutusuuringud vajalikud vaid siis, kui õigekirjareegel puudub või on ebajärjekindel. Artiklis tuuakse näitena välja juhtumiuuring ühest konkreetsest keelekasutusprobleemist andmebaasis, kirjeldades detailselt keelenõuannete väljatöötamiseks kasutatud strateegiat, keskendudes eriti korpusepäringute asjakohasele kasutamisele.

Märksõnad: keelenõustamine, korpused, korpuspäringud, keelekasutusprobleemid, õigekirjaprobleemid