Articles

A Data-Driven Analysis of Reader Viewpoints: Reconstructing the Historical Reader Using Keyword Analysis*

Masako Fidler and Václav Cvrček

Abstract: This study uses corpus-linguistic methods to examine the relationship between language usage patterns and divergence in text interpretation. Our target of analysis is a set of texts (Czechoslovak presidential New Year's addresses from 1975 to 1989), which contemporary readers consider repetitious and devoid of content. These texts were statistically contrasted with corpora from two different periods: one from the totalitarian period and the other from the contemporary (post-totalitarian) period. The comparison was based on the Difference Index, the most recent effect-size estimator, which was used to enhance the interpretation of keyword analysis outcomes. The two analyses yield significantly different results: the data from the analysis using the contemporary corpus were commensurate with contemporary readers' impressions; those from the analysis using the totalitarian corpus fluctuated in tandem with (and sometimes in anticipation of) political and social changes during the 15-year period and suggested an interpretation of the texts by a reader more familiar with totalitarian texts.

1. Introduction

The existing literature on discourse suggests that a set of expectations plays a crucial role in the interpretation of events, actions, and texts.[1]

* The authors would like to thank the two anonymous referees who provided valuable input. We would like to thank Andrew Malcovsky for careful copyediting of the manuscript. Responsibility for any errors in the resulting work remain our own. This project was partially funded by the Brown University Humanities Research Funds and was written under the auspices of the Programme for the Development of Fields of Study at Charles University, No. P11 Czech national corpus.

[1] This interaction between viewpoints and the target of linguistic investigation goes back to de Saussure: "Other sciences work with objects that are given in advance and that can then be considered from different viewpoints; but not linguistics. […] it would seem that it is the viewpoint that creates the object […]" (Saussure 1916/1959: 8)

Schank and Abelson (1977) use the term "script" to refer to an anticipated order of events, such as the structure of a seminar presentation or ordering a meal in a restaurant. A number of studies use the term "schema" in describing predictable narrative patterns (Bartlett 1932; Labov and Waletzky 1967; Labov 1972; Chafe 1986, 1994).[2]

Tannen and Wallat (1993) propose the concepts of interactive frames and knowledge schemas; the former concerns "what is going on" in each interaction (59), while the latter represent "knowledge structures," i.e., what people expect about individuals, entities, events, and settings (60). The latter are particularly important in considering how a given text is received. We assume that knowledge structures are closely intertwined with what a reader thinks s/he will find in a text, and that these structures directly impact reader interpretation. This is commensurate with a reader's pragmatic quest for optimal relevance in text interpretation, as argued by Sperber and Wilson (1986) and Blakemore (2003: 105). Readers are likely to process new information to yield an improvement to their representation of the world (e.g., confirmation or modification to what s/he already knows) with the minimum amount of effort—with the help of his/her extant salient knowledge structures.

When two individuals are exposed to markedly different types of cultural and social values, then it is reasonable to anticipate that each of them will find different topics more striking than others. According to Bakhtinian dialogism, individuals' interpretations of the world can never be the same because interpretations emerge in concrete social contexts as a result of "unique relations between the self, others, and the outer world" (Lähteenmäki 1998: 88). Tannen (1979) demonstrates this empirically, showing different interpretations of different points in the "pear film" by two viewers: speakers of English and speakers of Modern Greek (146–75).

The existing literature, in short, suggests that reception of a text can vary among readers with different expectations. We anticipate that such expectations are inevitably built on patterns of language use through consistent exposure. This idea is not entirely new; other scholars have made observations that point in this direction. Take, for example, the phrase *illegal immigrant*. The two words prime one another through repetition; readers exposed to this phrase are likely to think of this sequence even when encountering the word *immigrant* without

[2] The term schema covers a wider conceptual notion in cognitive linguistics (cf. Tuggy 2007).

a modifier; for such readers the word often invokes suspicion about migration (Stubbs 1996: 197). Language use thus leads to the reader's conceptualization of the world (point of view).

In this paper we present empirical evidence for the relationship between language use and the time-sensitive nature of discourse interpretation. We will do so by using a particular strand of corpus-assisted keyword extraction. Section 2 presents the goals of this study, which differs from related studies in the type of diachronic quantitative research pursued. A more detailed description of the advantages of this method can be found in Appendix 1. Section 3 presents the data, followed by their interpretation. The conclusions of our research and its further implications comprise section 4.

2. Methodology and Goals

2.1. Quantitative Approaches to Discourse Analysis

The study of discourse has been increasingly linked to quantitative analysis using language corpora. Statistical approaches help to reduce researcher bias and complement the qualitative analysis of texts (Baker 2004b: 346; for discussion see Baker 2006: 10). Raw corpus data and data subjected to statistical analysis are utilized in many areas of linguistics. Some studies investigate the discourse functions of grammatical categories (the historical present in Modern Greek [Thoma 2011] and innovation in Indian English [Sedlatschek 2009]), while another compares conceptual and stylistic patterns across national literatures (Jockers 2013). More recent discourse analysis often involves quantitative data, examining representations of individuals and society, e.g., Baker 2012 and David et al. 2013. A corpus approach to discourse is especially relevant since a text segment can be viewed as "a unique occurrence rather than a token" (Teubert 2005: 4). In other words, the meaning of a piece of text is embedded in a specific context and makes reference to a "unique set of other texts." Context in its widest sense constitutes societal and cultural knowledge, which is largely transmitted by language and is expected to fluctuate over time. We expect that context is closely connected with patterns of language use, and that these patterns are observable in diachronic language corpora—more specifically, that keywords extracted from a text with our method of keyword ranking

(Difference Index—henceforth DIN) reflect how a text is interpreted by a typical reader[3] of a specific period of time.[4]

2.2. The Study of Diachronic Text Reception

This section presents a brief description of the particular type of corpus linguistic keyword analysis performed in this study and compares our approach with the existing studies on diachronic text reception.

Any keyword analysis (KWA) involves a contrast between the target text (Ttxt) and a reference corpus (RefC) and yields a set of keywords (KWs). KWs are words that are statistically prominent in the Ttxt relative to their status in the RefC. For example, a KWA contrasting the fairy tale *Rusalka* (Water nymph) as a Ttxt against the background of a well-balanced RefC is expected to yield keywords such as *vodník* 'water goblin' and *ježibaba* 'witch', since the relative frequencies of these words are statistically more significant in the Ttxt than in the RefC, which reflects a much broader general language-usage pattern. In contrast, the same Ttxt compared to the background of RefC that consists of all Czech fairy tales is less likely to yield *rusalka* and *vodník* as KWs, as they are very frequently used in this particular genre. In this present study we utilize this "surprisal" aspect of KWAs, contrasting the same text with different RefCs.

Keywords can be extracted on multiple contextual levels:[5] e.g., a study of section-specific characteristics within an opus (e.g., a chapter as the Ttxt and the book containing that chapter as the RefC) and a study of author-specific characteristics within the entire language (all available texts written by one author as the Ttxt and a corpus that reflects general language patterns as the RefC). Some examples of section-specific studies include Sardinha 1996, 1999a, 1999b (comparison of a small corpus of business reports vs. RefC of 17 reports), Culpeper 2002 (comparison of the lines of specific characters vs. the lines

[3] The analysis, of course, anticipates that the "idealized" or prototypical reader who is exposed to a language pattern is reflected in the reference corpus to varying degrees.

[4] The data themselves are "keys" to text interpretations, "giv[ing] access to features of a text or corpus that are not immediately obvious" (Bondi 2010: 3). Researchers of course must process the data to arrive at their interpretations.

[5] Bertels and Speelman (2013: 554) only distinguish two levels of investigation, but in principle a comparison of texts between specific and general can be carried out with varying degrees of granularity.

of other characters in Shakespeare's plays), and Scott and Tribble 2006 (comparison of Romeo and Juliet vs. all Shakespeare plays). Examples of genre-specific studies include Baker 2004a (comparison of gay narrative texts with the British National Corpus) and Baker 2009 (comparison of transcripts from pro- and anti-fox-hunting debates in the British House of Commons with the Freiberg-Lancaster/Oslo-Bergen corpus).

Some studies extract keywords by comparing two different corpora. For example, Fairclough compares Tony Blair's "New Labor" texts (documents and texts from the media) and the "Early Labor" texts. This study is diachronic in that it examines changes in Labor Party ideology over time by comparing newer and earlier texts. A diachronic study by Baker (2010) looks at changes in the representations of Muslims and Islam between 1998 and 2008 in UK newspapers; it compares articles about Muslims and Islam in British tabloid newspapers (22 million words) with similar articles in major nationwide UK newspapers (65 million words).

The present KWA method shares its principles of KW extraction with other KWA methods but differs in several aspects. First, we use fifteen Ttxts, each of which is small; this allows close cross-examination of each text both qualitatively and quantitatively. Second, our Ttxts differ from those used in existing diachronic studies. Published in consecutive years from 1975 to 1989, they belong to the same genre; they are texts from the "normalization" period in Czechoslovakia after its socialist reform movement was crushed by the Warsaw Pact invasion of 1968. The properties of Ttxts are summarized in Table 1 below:

Table 1. Properties of Ttxts

	Size	**Period**	**Genre**
Ttxts	1,000–2000 words each; 22,088 words in total	1975 to 1989	New Year's Address by the socialist Czechoslovak President Gustáv Husák

Most importantly, the Ttxts are generally viewed as ritualistic and lacking real content.[6]

[6] Cf. Homolová, http://www.svet.czsk.net/clanky/publicistika/prezprojevy.html (accessed 27 March 2013) or an illustrative video at www.youtube.com/watch?v=QiBm4YX9y24 (accessed 8 October 2015), a pastiche of the New Year's Addresses from 1976 to 1989 to make one whole long sentence.

This apparent monotonousness of texts is highly symptomatic, if not unique, of Czech political speech from this period.[7] In contrast to existing KWA studies that **anticipate** changes in their Ttxts over time, the current study examines Ttxts that are generally assumed to be nearly **identical** and lacking relevant political messages.

Moreover, our study contrasts Ttxts with two RefCs from different times, one that reflects contemporary language use (SYN2010) representing all genres and another that consists of publications from the past, i.e., the totalitarian period (TOTALITA).

The two RefCs are different. SYN2010 represents all genres of contemporary language usage and is considered to be generally representative of the written language, whereas TOTALITA represents language patterns predominantly in socialist periodicals. The corpora, however, constitute a maximum contrast: SYN2010 approximates a reader who is hardly exposed to texts from the socialist period, whereas TOTALITA approximates a reader who was closely following the official press available during socialism.[8] The following table summarizes the properties of the two RefCs.

Table 2. Properties of RefCs

	TOTALITA	**SYN2010**
Size	12 million tokens	100 million tokens
Type of texts	corpus of journalistic and propagandistic texts from the 1950s through the 1970s in the former socialist Czechoslovakia.	a balanced synchronic representative corpus of written Czech (http://wiki.korpus.cz/doku.php/cnk:syn2010)

The use of two RefCs is not novel, but we use them for different goals than other studies. While the existing literature follows changes in the **production** of the text (i.e., characteristics of texts) over time, we

[7] Compare these texts with others, e.g., texts produced in Poland, where there were more outspoken protests against the regime, or texts from the USSR that played the leading role at different junctures of the Eastern-bloc history.

[8] TOTALITA is therefore a more artificially constructed "reader" than the contemporary counterpart based on SYN2010 because the former is predominantly based on one genre (periodicals).

examine **the relationship between the data and reader reception** from different times.[9]

Finally, our study applies the Difference Index (DIN) to rank KWs, which refines methods used and/or proposed in existing studies. The following section provides a concise definition of KWs, KWA, and a description of DIN.

2.3. Keyword Analysis

The identification of prominent words that play a potentially crucial role in text interpretation is normally the starting point of many empirical studies. This section will discuss methods of isolating prominent words with the help of quantitative methods based on the frequency with which elements appear in a text.

The major task of quantitative methods like these is to find words with "keyness." A word has keyness both when there is a statistically significant difference between its relative frequency (i.e., raw frequency divided by the size of the text) in the Ttxt and in the RefC and when its relative frequency in the Ttxt is higher than its relative frequency in the RefC. A word fulfilling both of these criteria is a KW and is connected with what the text is about and its stylistic characteristics (Scott 2010: 43).

The most commonly used tests to identify KWs are the chi-square or log-likelihood tests.[10] KWs in this sense therefore should not be confused with query or search-engine terms. They are not hand-picked words that carry specific associations and values within a community (Firth 1945: 40–41). KWs, as we use this term, also do not refer to "cultural keywords" that are associated with a culture and a society (Williams 1976) or words that facilitate the understanding of cultures and societies (Wierzbicka 1997, 2006, and 2010).

The process of KW identification is conceptually different from the method of finding words with thematic concentration (TC) (Popescu 2007 and Popescu et al. 2009), which is used in analysis of Czech polit-

[9] As this approach is expected to facilitate our understanding of how texts might be received by a model reader of a specific time, we intend to test this method to see the extent of its predictive power about reader reception of a text before it has even been published.

[10] Another suitable candidate, the Fisher exact test, is used less frequently, e.g., Bertels and Speelman 2013.

ical texts by David et al. (2013). As mentioned above, KWs are obtained by contrasting the Ttxt with a frame of reference (the RefC). The use of different RefCs can therefore result in different sets of KWs (cf. Scott and Tribble 2006 and Baker 2009, mentioned above in section 2.2). TC words, in contrast, are obtained from the study of one single text;[11] the set of TC words in a text is therefore invariant.[12]

Recent discourse studies based on large corpora have shown the importance of a ranking of KWs in which effect size plays a crucial role (cf. Appendix 1). Our KW-ranking method (DIN) is different from the traditional methods. The former ranks the KWs by effect size (which is derived from relative frequencies), whereas the latter uses statistical significance (test statistic value) based on raw frequencies. The details of the methodological advantages DIN over the currently existing methods are in Appendix 1.

As mentioned in section 2, we assume that the key to discourse interpretation is the reader's expectations. Our foundational assumption is that a RefC can approximate a model reader's exposure to language patterns, which in turn reflect the typical reader's point of view. This is a reasonable assumption, as noted by Taylor (2012), who discusses the relationship between existing language corpora and individuals' "mental corpus." Bermel et al. (2014) study the relationship between frequency and native speaker intuition in grammaticality judgment. They conclude that proportional frequency of forms is more closely associated with speakers' impressions and their linguistic behavior than absolute frequency. This observation parallels our assumption that raw frequencies in the Ttxt does not indicate prominence per se; words should be identified as key terms on the basis of differences between their relative frequencies in the Ttxt and RefC.

[11] In search for a word with TC, we must first find the "h-point" in the frequency list of word-types that occur in the text. The h-point is represented by a word with a frequency equal to its rank (e.g., the 57th word in the frequency list of types in a text with the raw frequency of 57 occurrences). This h-point splits the distribution into two different populations: words with a frequency higher than the h-point (usually grammatical words) and all other words (usually lexical/content words). TC words are those lexical or content words which can be found in the "grammatical part" of the distribution (i.e., above the h-point).

[12] TC has various other advantages and applications, e.g., comparing texts according to their thematic compactness (cf. David et al. 2013).

3. Data and Analysis

As described in section 2.2, this study extracts KWs from the same set of Ttxts against the background of two RefCs from two different time periods, one that reflects periodicals from the totalitarian period (TOTALITA) and the other that reflects contemporary language use (SYN2010). Each of these two RefCs is used as a static entity representing one whole period, remaining unchanged over time. The dynamic array of Ttxts forms a time series, with each text representing one year. As the genre and discourse situation are held constant, genre-specific KWs can be identified and separated from the other KWs. The parameters of these texts are made maximally constant in terms of the author (Gustáv Husák,[13] Czechoslovak President during the 1970s and 1980s) and the genre (Presidential New Year's Address [NYA]). In short, KWA will essentially identify words that deviate from the general patterns of language use reflected in each of the two RefCs.

The popular opinion among today's readers is that the NYAs are devoid of content, as they were presented during the political stagnation after the Warsaw Pact invasion of Czechoslovakia. The use of KWA in this paper is therefore different from other studies that anticipate changes in text properties prior to data extraction.

In order to see whether there are differences in the KWAs produced by contrasting the Ttxts with two RefCs, we first present an overview with the top 50 KWs (ranked by DIN) from the entirety of Husák's texts. We then examine the properties of KWs that are keyed more continually than others. Finally, we look at groups of related KWs and the fluctuation of their keyness. The results will demonstrate how KWs are felt to be unusual over time when Ttxts are contrasted against the background of TOTALITA and SYN2010. Stability in keyness would point to an interpretation confirming the contemporary readers' view that Husák's texts repeat the same information regardless of events in and outside Czechoslovakia. The results from the two KWAs will be simultaneously compared to the historical sequence of events in Eastern Europe and the USSR.

[13] It is quite likely that these texts were written by Husák himself rather than a speechwriter. This, however, is not relevant since the focus is on how KWs reflect reception of the same set of texts.

3.1. Top 50 KWs from the Entire Corpus of NYAs

The top 50 KWs in all of Husák's NYAs together are shown in Appendix 2.[14] The similarities and differences between the KWAs based on SYN2010 and TOTALITA (SYN-KWA and TOT-KWA) are summarized in Tables 3–6 in this section.

Table 3. Genre-related KWs among the top 50 KWs

	SYN-KWA	**TOT-KWA**
Genre-related KWs shared by both KWAs	*drazí* 'dear' *spoluobčané* 'fellow citizens' *zdravíme* 'we greet' *pozdravuji* 'I greet'	
Genre-related KWs not shared by both KWAs		*přeji* 'I wish' *novoroční* 'of the new year' *zdravím* 'I greet' *vážení* 'dear (lit. respected) *přátelé* 'friends'

Both KWAs find KWs that are genre-specific,[15] such as those referring to [the new] year and those that are part of typical address forms: 'dear', 'fellow citizens', 'we greet', and 'I send greetings'. There are, however, differences between the KWs extracted from SYN-KWA and from TOT-KWA. First, TOT-KWA lists a larger number of clearly genre-related KWs than SYN-KWA. Besides the four KWs mentioned above, TOT-KWA attributes keyness to an additional five: 'I wish', 'new (year)', 'I greet', 'dear' (lit. respected, pl), 'friends (address form)'.

Second, SYN-KWA gives higher ranking (i.e., sensitive) to more period-specific socialist KWs than TOT-KWA (Table 4).

[14] For this study we set the significance level for log-likelihood test at 0.01 to identify KWs. We set the KWA to list all KWs with at least 3 occurrences in the Ttxts (excluded were: prepositions, conjunctions, and numerals).

[15] "Clearly genre-specific KWs" constitute a minimum set of KWs that are anticipated in Presidential New Year's Addresses: address forms (to the nation/to the members of the society except those specific to the socialist regime such as comrades [male and female]), greetings (e.g., 'I/we greet'), expressions of wish (new year's wishes such as 'I/we wish'), and references to the new year ('new', 'year'). Selection of these words is therefore not arbitrary but is to a certain extent subjective.

Table 4. Ranking of Period-specific KWs

	KWs	KW ranking in SYN-KWA	KW ranking in TOT-KWA
Address forms	*soudružky* '[female] comrades'	2	Not among the top 50 KWs
	spoluobčané 'fellow citizens'	11	1
	přátelé 'friends'	Not among the top 50 KWs	39
Adjectives	*soudružské* 'of a comrade, appropriate as a comrade'	5	Not among the top 50 KWs
	*bratrsk** [16] 'brotherly'	*bratrskému* [17] (6) *bratrskými* (12) *bratrských* (24) *bratrský* (34)	
	*osvobozeneck** 'liberating'	*osvobozeneckého* (3) *osvobozenecký* (9)	
	horečného 'feverish'	8	
	*vědeckotechnick** 'scientific-technological' [18]	*vědeckotechnické* (16) *vědeckotechnického* (17)	
	*socialistick** 'socialist'	*socialistického* (32) *socialistickými* (33) *socialistických* (42)	
	imperialistické 'imperialist'	25	

SYN-KWA ranks '[female] comrades', a socialist address form for women (as part of *soudružky a soudruzi* 'female and male comrades') the second highest, whereas TOT-KWA does not include this word form among the top 50 KWs. Obviously, this is due to the difference in the

[16] The asterisk (*) indicates that word forms with more than one inflectional morpheme are listed among the 50 KWs.

[17] KWs were extracted in (inflected) word forms rather than in lemmas, as grammatical information for inflected languages can be key to interpreting how KWs function. For example, a KW such as *fronty* 'front' is likely to be part of *národní fronty* 'the National Front, gen sg' and is unlikely to be a major participant in an event. In contrast, *fronta*, in the nom sg and the syntactic subject of a sentence, is likely to represent an entity that plays a more important role in an event than the gen form.

[18] Compare the instances per million (ipm) for this adjective in both RefCs: 227 ipm in TOTALITA versus 0.87 ipm in SYN2010.

distribution of socialist terms in the two RefCs: TOTALITA contains many instances of "comrade" address forms, whereas SYN2010 rarely uses them; the former do not find these words as "surprising" as the latter.[19] Conversely, TOT-KWA gives higher ranking (the highest) to 'fellow citizens,' which is not an automatic socialist address form, than SYN-KWA; TOT-KWA also yields 'friends', a neutral address form, among the 50 KWs.

Period-specific adjectival forms and nominal forms are also ranked high in SYN-KWA, whereas they are not included in the top 50 in TOT-KWA. Adjectives that are used frequently during the socialist period are highly ranked in SYN-KWA: 'of comrade/appropriate as a comrade', 'brotherly', 'liberating', 'feverish', 'scientific-technological', 'socialist', and 'imperialist'.

The adjective *bratrsk** was associated with the USSR and other socialist countries (*bratrská spolupráce se Sovětským svazem* 'brotherly collaboration with the Soviet Union'). The word *osvobozeneck** usually occurs in the context of the struggle against oppression and imperialism (*počátek nové epohy osvobozeneckého boje mezinárodního proletariátu* 'The beginning of a new epoch of the liberating struggle of the international proletariat'). The word *vědeckotechnick** is mostly connected with the notion of progress during the socialist period (*vědeckotechnický pokrok/rozvoj* 'scientific-technological progress/development). The form *imperialistické* is a frequent adjective form to refer to western-bloc countries and their actions, especially the USA (*CIA je nadále nástrojem americké imperialistické politiky a používá stejné metody, […]* 'The CIA continues to be the instrument of American imperialist politics and uses the same methods, […].'). Word forms referring to the socialist five-year plans (*pětiletky* and *pětiletce)* and imperialism (of capitalist countries) (*imperialismu*) are only included only among SYN-KWA's top 50 KWs (as these words are either missing from contemporary language use or dying out).

Third, SYN-KWA and TOT-KWA diverge in their sensitivity to grammatical person and number in finite verb forms (Table 5);[20] the latter ranks the 1st person singular forms much higher. The genre-specific verb forms of the lemma *pozdravovat* 'to relay greetings', *zdravit* 'to greet', and *přát* 'to wish' are among the KWs in both KWAs. The two

[19] These words are marked in present-day Czech; they are used ironically or as a citation.

[20] Only nonpast (present and future) forms are relevant. Past-tense forms do not indicate person.

KWAs, however, differ in ranking of the inflected forms. TOT-KWA ranks the 1st pers sg forms *pozdravuji*, *zdravím*, and *přeji* much higher than SYN-KWA: *pozdravuji* is ranked much higher in TOT-KWA than in SYN-KWA (7th vs. 31th); *zdravím* and *přeji* are not even among the top 50 KWs in SYN-KWA. The data from TOT-KWA suggest a perception that the speaker is presenting himself as an individual more than the SYN-KWA.[21]

Table 5. 1sg verb forms

Grammatical forms in 1sg	Ranking in TOT-KWA	Ranking in SYN-KWA
pozdravuji 'I relay greetings to'	7	31
zdravím 'I greet'	19	176
přeji 'I wish'	12	117

Fourth, SYN-KWA suggests reception of the Ttxts as descriptive. Its prominent KWs present the Ttxts as static rather than dynamic; it is adjectives rather than other parts of speeches that are highly keyed. The popular reception of Ttxts as static is commensurate with the type of KWs obtained from SYN-KWA. The distribution of adjectival forms as opposed to the others is statistically significant (Table 6).

Table 6. Adjectival forms among top 50 KWs from TOT- and SYN-KWAs

	Top 50 KWs: Adjectival Forms	Top 50 KWs: Others	Total
TOT-KWA	11	39	50
SYN-KWA	26	24	50
total	37	63	100

Chi-square statistic 9.6525; $p < 0.005$

[21] This observation follows from the data obtained and from our assumption that TOTALITA approximates a model reader from the socialist period. However, we acknowledge that prominence of these word forms may have resulted from the properties of TOTALITA (socialist periodicals and journalistic texts) where expressions of personal interaction are rarer.

KWs from SYN-KWA also suggest that the notion of collectivity (expressed by the 1st pers pl form *zdravíme*) might be viewed prominently. In contrast, TOT-KWA KWs draw attention to actions rather than description with a predominance of verb forms among the top 50 KWs: 13 verb forms from SYN-KWA and 22 forms from TOT-KWA are among the top 50 KWs.[22]

3.2. Stability in Keyness

The KWs from SYN- and TOT-KWAs differ in the way they manifest keyness over the 15-year period (Appendix 3). In general, there are more than double the number of the same KWs that are continually keyed (i.e., repeatedly from year to year) in SYN-KWA compared to TOT-KWA. This, however, does not by itself mean that the texts are viewed as more **informative over time** in SYN-KWA than in TOT-KWA. On the contrary, as the same set of KWs from each NYA are given the same weight from year to year, each text yields a similar interpretation. This is consistent with the popular impression among present-day readers of NYAs that these texts are "repetitious and uninteresting." In contrast, the smaller number of continual KWs in TOT-KWA suggests that different word forms are keyed from year to year. KWs from TOT-KWA, in comparison to those from SYN-KWA, indicate sensitivity to subtle changes in text reception over time, which are connected with the perception of ongoing and upcoming political changes. Also, as the TOT-KWA KWs are varied and the vast majority of them are present only in one or two NYAs, they show that Husák's NYAs are far from repetitive from the viewpoint of language use in the socialist period.[23]

In the following section we will examine groups of semantically related KWs and compare their keyness over time in order to demonstrate

[22] These differences may also suggest language change since the 1970s. Contemporary language is more dynamic and individualistic and contains fewer formulaic expressions, not to mention socialist terminology.

[23] It is well documented that the number of KWs identified in each text is influenced by the size of the RefC (Scott and Tribble 2006: 64). Since SYN2010 is almost 10 times larger than TOTALITA, the number of recurring types may be influenced by the inequality of RefC sizes. The more KWs detected in each year, the higher the probability of KWs recurring in different NYAs. However, given the overall number of KWs in each text (in comparison to the number of KWs with continual keyness), we came to the conclusion that the RefC size has negligible effect on our findings.

that the results from TOT-KWA more closely match political changes in the country than those of SYN-KWA.

3.3. Keywords in Semantically Related Groups

The keyness of related KWs shows more dynamic fluctuation in TOT-KWA than in SYN-KWA.[24] This is illustrated by three groups of KWs labeled "Cold War," "Collective Possession," and "Ideological Markers" that were keyed at least once in each of the KWAs from 1975 to 1989.[25] Table 7 on page 212 shows the list of KWs that were grouped together. The justification for grouping the word forms, the extracted data, and their interpretation follows.

3.3.1. Cold War KWs

Cold War KWs belong to one of the predominant topics in socialist discourse. They were frequently used in Husák's NYAs to present the capitalist world as encroaching on the peace-loving socialist nations.

(1) Na mezinárodním poli nejreakčnější imperialistické kruhy ve snaze udržet své otřesené pozice nastoupily kurs na zostřování **napětí** a vyvolávání konfliktů a konfrontací v různých částech světa. Kladou nejrůznější překážky politice míru, **mírového** soužití a uvolňování **napětí**. (1981)

'On the international arena the most reactionary imperialist circles, in an effort to maintain their shaken positions, launched a course [of action] for the sharpening of **tension** and the provocation of conflicts and confrontations in various parts of the world. They place the most diverse obstacles to the policies **of peace**, **of peaceful** coexistence, and release **of tension**.'

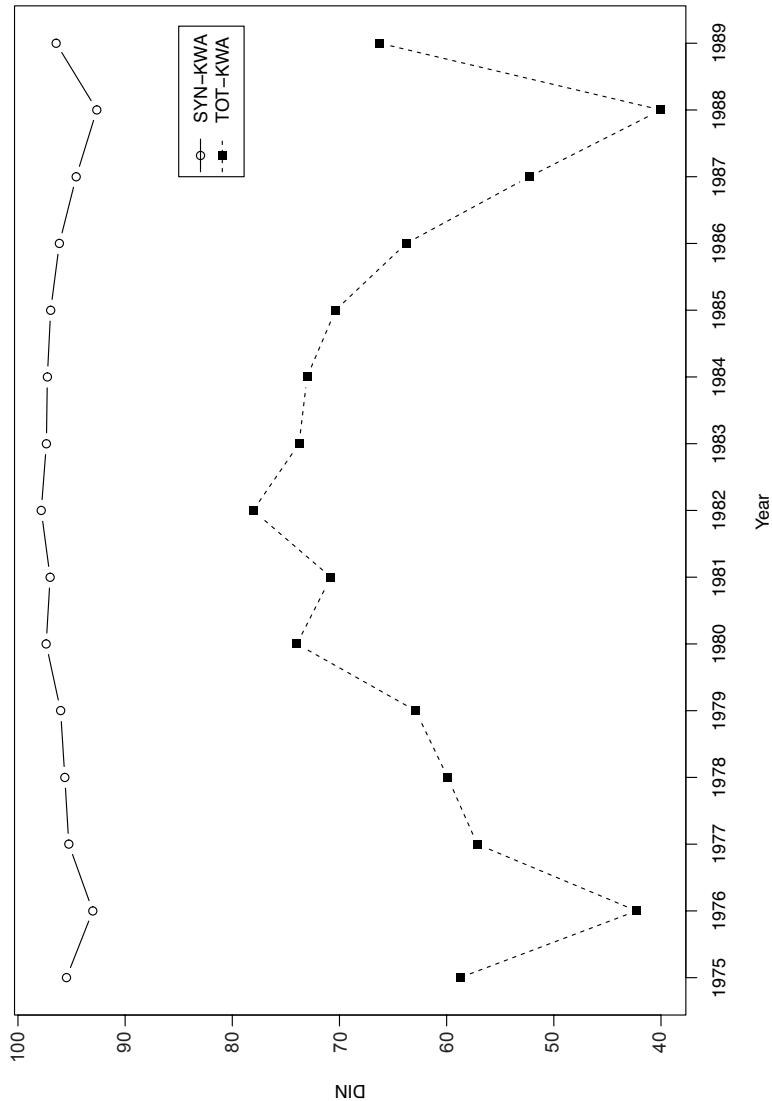
As shown in Graph 1 on page 213, Cold War KWs maintain the same level of keyness throughout the 15-year period in SYN-KWA. Keyness of these words in TOT-KWA, in contrast, fluctuates, the fluctuation co-

[24] The quantitative data can be obtained on request from either of the authors.

[25] There is admittedly a degree of subjectivity in these groupings, but such a procedure is widely used in keyword analyses (cf. Baker 2009).

Table 7. Semantically Related KW groups

Cold War KWs	
a. KWs connected with peace	Nominal form 'peace': *mír, míru*, Adjectival forms 'of peace': *mírová, mírové, mírového, mírovému, mírovou, mírový, mírových, mírovými* Adjectival forms 'peace-loving': *mírumilovné, mírumilovných, mírumilovným*
b. KWs connected with tension	Nominal form 'tension': *napětí*
c. KWs connected with (dis)armament	Nominal forms 'disarmament': *odzbrojení* Nominal form 'weaponry': *výzbroje*, Nominal forms 'arms': *zbrojení, zbrojením* Participial form 'carrying out disarmament' *odzbrojovací* Adjectival forms 'armed': *ozbrojených*,
Collective Possession	
inflected forms of the 1st pl possessive pronoun 'our'	*náš, naše, našeho, našem, našemu, naši, naší, našich, našim, naším, našimi*
Ideological Markers	
a. KWs connected with socialism	Nominal forms 'socialism': *socialismu, socialismus* Adjectival forms 'of socialism': *socialistická, socialistické, socialistického, socialistickém, socialistickému, socialistickou, socialistický, socialistických, socialistickým, socialistickými*
b. KWs connected with communism	Nominal form 'communism': *komunismu* Nominal form 'of communism': *komunisté, komunistů*, Abbreviated nominal form 'Czechoslovak Communist Party': *ksč* Nominal forms 'communist': '*komunistům, komunisty, komunistická, komunistické, komunistickým*,



Graph 1. Keyness of the Cold War KWs in SYN- and TOT-KWA

inciding with domestic political shifts as well as external changes in the Eastern Bloc nations.[26]

Keyness rises earlier, and peaks in the early 1980s, when the tension in Central Europe heightens, owing to demonstrations, subsequent martial law in neighboring Poland, and the US sanctions against that country. Keyness begins to decline as the political situation seemingly stabilizes on Gorbachev's rise to power in the USSR in 1985 and the commencement of disarmament negotiations, only to rise again in 1989 when outspoken protests against the government in Slovakia and in the Czech lands spread.

3.3.2. Collective Possession

Various forms of the 1st person plural possessive pronoun are frequently used in socialist discourse. The pronoun presents events as collective rather than individual actions. It also implicitly distinguishes one group of people ("us") from the other ("them").

(2) V loňském roce se uskutečnily rovněž všeobecné volby do zastupitelských sborů, ve kterých se demokraticky obnovily orgány státní moci, dále se upevnil **náš** socialistický stát. Výsledky voleb se staly velkým politickým vítězstvím **našeho** lidu. (1982)

'In the last year there also took place general elections for the representative bodies, where the organs of state power were democratically renewed, and **our** socialist state was further solidified. The results of the elections became a great political victory of **our** people.'

An individual is expected to support the socialist state in order to qualify as one of "our people." The keyness of these possessive pronouns also shows more explicit fluctuation in TOT-KWA than in SYN-KWA.

Collective Possession KWs show little change in keyness during the 15-year period in SYN-KWA. TOT-KWA, on the contrary, shows visible decline from 1977 to 1989. It is worth noting that Collective Possession

[26] The KWs extracted from both analyses are similar, with a shorter list from TOT-KWA. The differences between the analyses reside in the prominence of KWs, in other words, the extent to which KWs deviate from the language-usage patterns reflected in the respective RefCs.

declines in keyness from 1985 to 1988 even after the Soviet leadership stabilized, an indication that it was becoming increasingly difficult for the Czechoslovak government to speak in the name of its people or describe its actions as having their sanction. The very low keyness of Collective Possession and the sudden blip in keyness of Cold War KWs in 1989 (see Graph 2 on page 216) correspond to the period when the political leadership was making a futile attempt to justify the status quo with Cold War rhetoric but was unable to negate an intensifying popular disconnect.

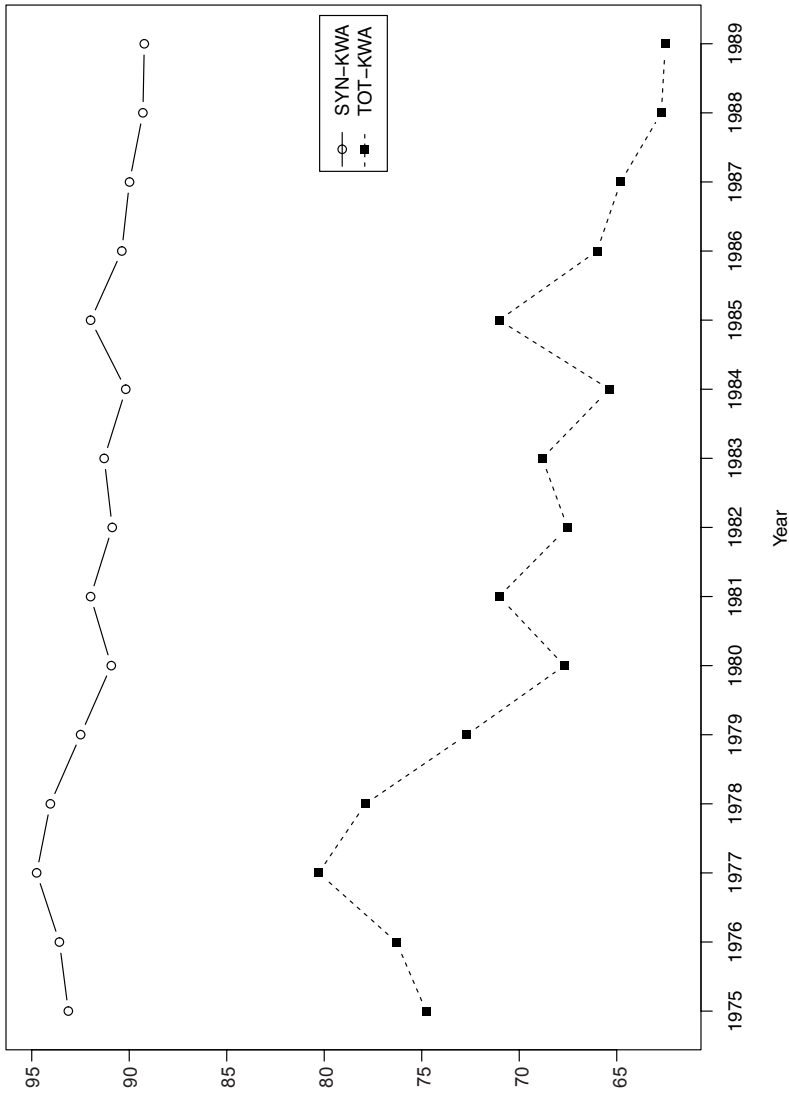
3.3.3. Ideological Markers

This category consists of ideological adjectives that carve out ideas, entities, and individuals from more general categories. Socialist democracy is not simply democracy; it is a special kind of democracy that is different from other possible forms of democracy.[27] Socialist homeland is not merely one's homeland, but a homeland under socialism.

(3) V činorodé práci se prohlubovalo **socialistické** uvědomění, posilovala se jednota tříd a sociálních vrstev. Utužoval se bratrský svazek Čechů, Slováků a národností našeho státu, upevnilo se naše **socialistické** zřízení. Rozvíjelo se spojení vedoucí síly naší společnosti, **Komunistické** strany Československa, s nejširšími vrstvami lidu. Výrazem předností naší **socialistické** demokracie je široká účast pracujících na správě a řízení státu, v činnosti Národní fronty, národních výborů, jakož i celková jejich tvořivá iniciativa a občanské angažovanost. (1979)

'Through effective work, **socialist** awareness deepened, unity of classes and social strata strengthened. The brotherly bond of Czechs, Slovaks, and the ethnic groups of our state became firm, our **socialist** system solidified. There developed a union of the leading power of our society, the **Communist** Party of Czechoslovakia, and the widest strata of the people. An expression of the superiority of our **socialist** democracy is the broad participation of workers in the administration and management of the state, in the activity of the National

[27] A similar observation is made by Andrews (2011: 2) on totalitarian languages.



Graph 2. Keyness of Collective Possession in SYN- and TOT-KWA

Front, the national committees, as well as their entire creative initiative and civic engagement.'

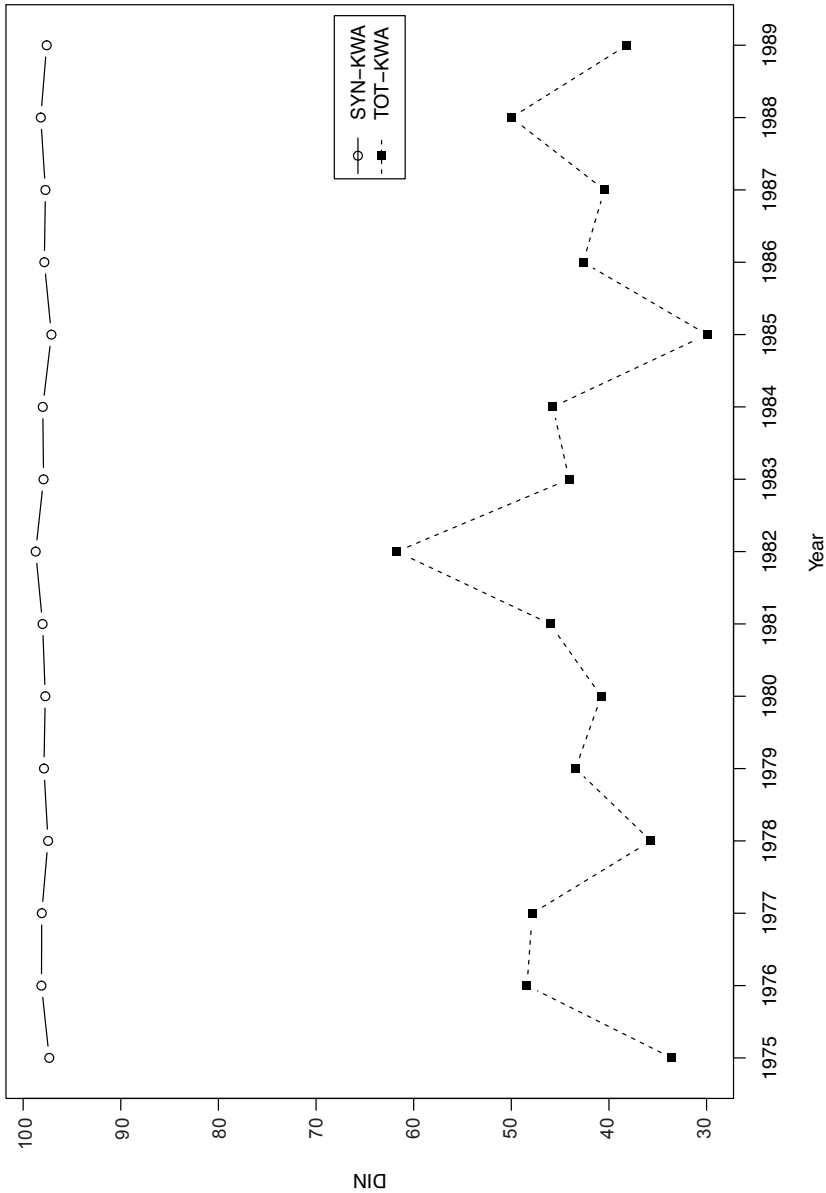
Thus, the adjective *socialist* specifies that active work leads to a deepening of a socialist awareness and the socialist system. The leading power of society is assumed to belong to the communist party (rather than any political party in power at the time). People are said to express preference for socialist democracy (rather than for democracy in general).

The keyness of Ideological markers remains stable in SYN-KWA throughout the time period. In contrast, the same group of KWs shows different degrees of keyness over time in TOT-KWA as well. See Graph 3 on page 218.

The rise and fall of keyness in this graph coincides with key events that triggered a need to maintain the status quo of the socialist regime. The drop from 1982 to 1983 and further into 1985 is seen in the period when the USSR cycled through three General Secretaries shortly after Brezhnev's death (Andropov 1982–84, Chernenko 1984–85, and Gorbachev 1985). The rise of keyness of the ideological markers at the beginning and end of the study period coincides with a need to align with the rest of the socialist bloc nations during the demonstrations in Poland in the early 1980s and Gorbachev's perestroika period during the second half of the 1980s.

3.3.4. Consistent Differences between KWAs with Respect to Semantically Related KWs

As seen throughout sections 3.3.1 to 3.3.3, semantically related KWs from TOT-KWA show more visible fluctuations in keyness than those from SYN-KWA. Such fluctuations occur in tandem with or immediately preceding major political and societal changes. The KWs from SYN-KWA maintain nearly an identical level of keyness; the level of keyness is in general high. This is in agreement with the impression voiced by contemporary readers that Husák's speeches are full of socialist jargon (which is outdated and therefore unusual) but are "all the same" each year. The much lower prominence of KWs in TOT-KWA indicates a different reading of the same texts. The mere occurrence of socialist jargon is not surprising (thus the much lower prominence of KWs). The sharp contrast to totalitarian discourse instead helps distin-



Graph 3: Ideological markers in SYN- and TOT-KWA

guish the fine-grained fluctuations of socialist clichés, which are connected with political and societal changes.

4. Concluding Observations

This paper has shown that changing the reference corpus in KWA significantly changes the interpretation of the same text. The results of this study show consistent differences between SYN-KWA and TOT-KWA. The prominent KWs from SYN-KWA present the Ttxts as more static rather than dynamic; adjectives rather than other parts of speech are highly keyed. SYN-KWA also does not attribute prominence to use of the 1st pers verb forms (*zdravím* 'I greet' and *přeji* 'I wish') as TOT-KWA (Table 3); the action of the speaker is thus not as highlighted by the former as the latter. These observations are commensurate with the contemporary reader's perception of Husák's NYAs that they are repetitious and ritualistic.

Each KWA ranks different types of KWs as more prominent: SYN-KWA, which represents the model reader who is exposed to the contemporary language-usage patterns, lists more KWs specific to the socialist period than TOT-KWA. The former lists a more constant set of "socialist KWs" than the latter and attributes nearly equal high keyness over the entire period of 15 years compared to the three groups of KWs that represent the socialist period. These results correspond to the impression of today's readers that Husák's NYAs repeatedly use socialist clichés and do not address the burning issues of the time. In other words, the socialist clichés are so prominent and unexpected that it distracts the reader's attention, not allowing the reader to see subtle changes that might be indicative of political and societal shifts.

In contrast, TOT-KWA, which represents the reader from the past (i.e., the model reader who is highly exposed to the official propaganda discourse), lists fewer adjectival forms among the top 50 KWs than SYN-KWA. The former does not constantly attribute the same degree of keyness to the same KWs; keyness of the three semantic KW groups fluctuates at different points in history. KWs from TOT-KWA suggest a more dynamic impression of the NYAs over the whole of the 15-year period.

Against the background of the historical events during the time the NYAs were made public, it is possible to conclude that TOT-KWA manifests more sensitivity to those events. It suggests the view of the

historical reader who could filter out the expected "fillers" and discern subtly prominent signs of change in politics and society.

In short, the data from this study show that KWA can reflect time-sensitive reception of the same text by using RefCs from different periods. A RefC from a specific period reflects patterns of language use, which in turn are connected with an overarching view of the model reader of that time. Keyword analysis with alternating RefCs can therefore serve as an initial step in constructing or reconstructing reader framing or reader expectations, concepts that have been discussed in discourse analysis.

It is highly possible that keyword analysis using the DIN can help predict how a text is likely to be received even before it is presented.

The present study is informative of the nature of Husak's texts: we can conclude that—contrary to the popular view of present-day readers—his texts might have been sending subtle messages to readers well versed in socialist discourse. Readers who are not experienced in socialist discourse are less likely to notice them.

Our approach differs from the existing literature on keyword analysis and reader expectations in discourse analysis. Application of KWA to Ttxts that are from the onset "uninformative," such as Husák's speeches, is unusual, as the method is mostly used when researchers expect positive characteristics from the onset. Our focus is on the reception of text rather than on its production (unlike, e.g., finding a common denominator in discourse, such as different representations of fox-hunting, as in Baker 2009). We used DIN-based ranking of keywords and shifts in keyword prominence against the background of two reference corpora to provide substantial empirical evidence to the interconnectedness between language use and varying interpretations.

The present paper, as the first detailed attempt at KWA of totalitarian speeches in Czech, points to further investigation in at least three areas. Interpretation of KWs requires steps (e.g., semantic grouping) that involve subjectivity; further reduction of such subjectivity is a task for future research. The present analysis implicitly treats continuous KWs as having the same semantic value. This problem will be pursued using KW links in our future research. Study of reader reception in other genres awaits further study.

References

Andrews, Ernest. (2011) "Introduction". Ernest Andrews, ed. *Legacies of totalitarian language in the discourse culture of the post-totalitarian era: The case of Eastern Europe, Russia, and China*. Plymouth, UK: Lexington Books, 1–13.

Baker, Paul. (2004a) "'Unnatural acts': Discourses of homosexuality within the House of Lords debates on gay male law reform". *Journal of sociolinguistics* 8(1): 88–106

———. (2004b) "Querying keywords: Questions of difference, frequency, and sense in keywords analysis". *Journal of English linguistics* 32: 346–59.

———. (2006) *Using corpora in discourse analysis*. London: Continuum.

———. (2009) "The question is, how cruel is it? Keywords in debates on fox hunting in the British House of Commons". Dawn Archer, ed. *What's in a word-list?* London: Ashgate, 125–36.

———. (2012) "Acceptable bias? Using corpus linguistics with critical discourse analysis". *Critical discourse studies* 9(3): 247–56.

Bartlett, Frederic. C. (1932) *Remembering: A study in experimental and social psychology*. Cambridge: Cambridge University Press.

Bermel, N., L. Knittl, and J. Russel. (2014) "Absolutní a proprocionální frekvence v ČNK ve světle výzkumu morfosyntaktické variace v češtině". *Naše řeč* 97(4–5): 216–27.

Bertels, Ann and Dirk Speelman. (2013) "'Keywords method' versus 'Calcul des Spécificités'". *International journal of corpus lingustics* 18(4): 536–60.

Blakemore, Diane. (2003) "Discourse and relevance theory". Deborah Schiffrin, Deborah Tannen, and Heidi E. Hamilton, eds. *The handbook of discourse analysis*. Malden, MA: Blackwell, 100–18.

Bondi, Marina. (2010) "Perspectives on keywords and keyness: An introduction". Marina Bondi and Mike Scott, eds. *Keyness in texts*. Amsterdam: Benjamins, 1–18.

Chafe, Wallace. (1986) "Beyond Bartlett: Narratives and remembering". Elisabeth Gülich and Uta M. Quasthoff, eds. *Narrative analysis: An interdisciplinary dialogue*. Special Issue of *Poetics* 15: 139–51.

———. (1994) *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. Chicago: University of Chicago Press.

Culpeper, Jonathan. (2002) "Computers, language and characterisation: An analysis of six characters in *Romeo and Juliet*". U. Melander Marttala, C. Ostman and Merja Kyto, eds. *Conversation in life and in literature: Papers from the ASLA Symposium, Association Suédoise de Linguistique Appliquée (ASLA) 15*. Uppsala: Universitetstryckeriet, 11–30.

Cvrček, Václav, and Masako Fidler. (2013) "Not all keywords are created equal: How can we measure keyness?" Paper presented at the Corpus Linguistics Conference, Lancaster, UK, July 2013. Available at https://docs.google.com/a/brown.edu/file/d/0B2ITZWv1AVaMalRKUW1pWWgyUWs/edit.

de Saussure, Ferdinand. (1916/1959) *Course in general linguistics*. Charles Bally and Albert Sechehaye, eds. Wade Baskin, trans. New York: Philosophical Library.

David, Jaroslav, Radek Čech, Jana Davidová Glogarová, Lucie Radková, and Hana Šústková. (2013) *Slovo a text v historickém kontextu*. Ostrava: Host.

Davies, Bronwyn. (2000) *A body of writing, 1990–1999*. Oxford: Rowman and Littlefield.

Fairclough, Norman. (2000) *New labour, new language?* London: Routledge.

Firth, John R. (1935) "Technique of semantics". *Transactions of the philological society* 34: 36–73.

Gabrielatos, Costas and Anna Marchi. (2012) "Keyness: Appropriate metrics and practical issues". Paper presented at the CADS International Conference, Bologna, Italy, September 2012. Available at http://www.gabrielatos.com/Presentations.htm.

Goffman, Erving. (1974) *Frame analysis: An essay on the organization of experience*. Cambridge, MA: Harvard University Press.

Hofland, Knut and Stig Johansson. (1982) *Word frequencies in British and American English*. Bergen: Norwegian computing centre for the Humanities.

Homolová, Marie. (n.d.) "Prezidenti, mistři novoročních přání". Available at http://www.svet.czsk.net/clanky/publicistika/prezprojevy.html, accessed 27 March 2013.

Jockers, Matthew L. (2013) *Macroanalysis: Digital methods and literary history*. Urbana: University of Illinois Press.

Kilgarriff, Adam. (2009) "Simple maths for keywords proc". Corpus Linguistics, Liverpool, UK, July 2013. Available at http://ucrel.lancs.ac.uk/publications/cl2009/171_FullPaper.doc.

Labov, William. (1972) "The transformation of experience in narrative syntax". William Labov, ed. *Language in the inner city: Studies in the Black English vernacular*. Philadelphia: University of Pennsylvania Press, 354–96.

Labov, William and Joshua Waletzky. (1967) "Narrative analysis: Oral versions of personal experience". June Helm, ed. *Essays on the verbal and visual arts: Proceedings of the 1966 Annual Spring Meeting of the American Ethnological Society*. Seattle: University of Washington Press, 12–44.

Lähteenmäki, Mika. (1998) "On meaning and understanding: A dialogical approach". *Dialogism* 1: 74–91.

Popescu, Ioan-Iovițˌ ed. (2009) *Word frequency studies*. Berlin: Mouton de Gruyter.

Popescu, Ioan-Iovițˌ Karl-Heinz Best, and Gabriel Altmann. (2007) "On the dynamics of word classes in texts". *Glottometrics* 14: 58–71.

Sardinha, Berber A. (1999b) "Using keywords in text analysis: Practical aspects". *DIRECT Papers* 42: 1–8.

———. (1996) "Review: WordSmith tools." *Computers and texts* 12: 19–21.

———. (1999a) "Word sets, keywords, and the contents: An investigation of text topic on the computer". *DELTA* 15(1): 141–49.

Scott, Mike. (1997) "PC analysis of key words—and key key words". *System* 25(2): 233–45.

———. (1999) *WordSmith tools help manual, Version 3.0.* Oxford: Oxford University Press.

———. (2010) "Problems in investigating keyness, or cleansing the undergrowth and marking out tails…". Marina Bondi and Mike Scott, eds. *Keyness in texts*. Amsterdam: Benjamins, 43–57.

Scott, Mike and Christopher Tribble. (2006) *Textual patterns: Keyword and corpus analysis in language education*. Amsterdam: Benjamins.

Sedlatschek, Andreas. (2009) *Contemporary Indian English*. Amsterdam: John Benjamins.

Shank, Roger C. and Robert P. Abelson. (1977) *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Hillsdale, NJ: Erlbaum.

Sperber, Dan and Deirdre Wilson. (1986) *Relevance: Communication and cognition*. Cambridge, MA: Harvard University Press.

Stubbs, Michael. (2001/2003) "Computer-assisted text and corpus analysis: Lexical cohesion and communicative competence". Deborah Schiffrin, Deborah Tannen and Heidi E. Hamilton, eds. *The handbook of discourse analysis*. Oxford: Blackwell, 304–20.

Tannen, Deborah. (1979) "What's in frame?" Roye Freedle, ed. *New directions in discourse*. Harwood, NJ: Ablex, 137–81.

Tannen, Deborah and Cynthia Wallat. (1987) "Interactive frames and knowledge schemas in interaction: Examples from a medical examination/interview". *Social psychology quarterly* 50(2): 205–16.

Taylor, John. (2012) *Mental corpus: How language is represented in the mind*. Oxford: Oxford University Press.

Teubert, Wolfgang. (2005) "My version of corpus linguistics". *International journal of corpus linguistics* 10(1): 1–13.

Thoma, Chrystalla A. (2011) "The function of the historical present tense: Evidence from Modern Greek". *Journal of pragmatics* 43: 2373–91.

Tuggy, David. (2007) "Schematicity". Dirk Geerarts and Herbert Cuyckens, eds. *The Oxford handbook of cognitive linguistics*. Oxford: Oxford University Press, 82–115.

TVhistorie. "Gustáv Husák—Novoroční projevy 1976–1989—Unikátní a zábavný sestřih". Available at http://www.youtube.com/watch?v=QiBm4YX9y24, accessed 27 March 2013.

Wierzbicka, Anna. (1997) *Understanding cultures through their keywords: English, Russian, Polish, German, and Japanese*. Oxford: Oxford University Press.

———. (2006) *English: Meaning and culture*. Oxford: Oxford University Press.

Wierzbicka, Anna. (2010) *Experience, evidence, and sense: The hidden cultural legacy of English*. Oxford: Oxford University Press.

Williams, Raymond. (1976/85) *Keywords: A vocabulary of culture and society*. New York: Oxford University Press.

Corpora Used

Křen, M, T. Bartoň, V. Cvrček, M. Hnátková, T. Jelínek, J. Kocek, J., R. Novotná. V. Petkevič, P. Procházka, V. Schmiedtová, and H. Skoumalová. (2010) *SYN2010: Genre-balanced Corpus of Written Czech.* Ús-

tav Českého národního korpusu FF UK Prague. Available at: http://www.korpus.cz.

Skoumalová, H, T. Bartoň, V. Cvrček, M. Hnátková, J. Kocek. (2010) *TOTALITA*. Ústav Českého národního korpusu FF UK, Prague. Available at: http://www.korpus.cz.

Masako Fidler
20 Manning Walk, Box E
Department of Slavic Languages
Brown University
Providence, RI 02912, USA
masako_fidler@brown.edu

Václav Cvrček
Institute of the Czech National Corpus
Faculty of Arts
Charles University in Prague
nám. J. Palacha 2
Prague 1, 116 38
Czech Republic
vaclav.cvrcek@ff.cuni.cz

Received: March 2015
Revised: July 2015

Appendix 1: Difference Index (DIN)

This study uses Difference Index (DIN) for ranking of KWs. This ranking method is different from the most frequent approaches based on statistical significance and from other effect-size estimators. DIN differs from the latter in tackling a situation where a RefC does not contain a word that occurs in a Ttxt. As DIN is not widely known yet, a brief description of the method and the theoretical motivation for using it is necessary.

The most widespread version of keyword analysis (Scott 2006) uses statistical significance, or *p*-values, to rank KWs. This method consists of three steps. First, the Ttxt is tokenized and the frequency of each word-type is counted. Second, the frequency of each word in the Ttxt is compared to the frequency of the same word in the RefC using the log-likelihood test, chi-square test, or Fisher's test. Finally, KWs are sorted **according to the value of the test**, which, in the case of log-likelihood or chi-square, can easily be converted to the *p*-value.

Ranking of KWs obviously plays a crucial role in the interpretation of a text, especially when the number of KWs is high. When the number of KWs exceeds a certain limit (e.g., thousands of KWs), it is then virtually impossible to examine each one of them carefully. As a result, researchers often try to reduce the number of KWs in a reasonably objective way. One solution is to drop the significance level (to 0.0001 or even less) so that fewer words qualify as having a statistically significant difference in their relative frequencies between the Ttxt and RefC. This has been proven inadequate (Gabrielatos and Marchi 2012; Cvrček and Fidler 2013),[28] as the *p*-value of a test represents only statistical significance and does not take into account the effect size (or the relevance) of the difference.

Another approach to deal with a large number of KWs is to pick only some of them (e.g., the top 100 or 1000). The *p*-value, however, does not provide reliable ranking of KWs necessary for this purpose. It might instead lead to a paradox where less prominent words might score higher than more prominent ones. This situation arises because

[28] Cvrček and Fidler (2013) compare the KW ranking from Husák's New Year's Addresses based on the *p*-value (log-likelihood test) and the KW ranking based on the effect size of KWs.

significance of frequencies does not by itself inform us of whether the difference between the frequencies in the Ttxt and the RefC carries any descriptive value. It only reveals **whether we have a sufficient amount of data** to conclude that the relative frequencies do in fact differ (i.e., they are drawn from two distinct populations). The larger the amount of data, the higher the likelihood that the resulting difference is significant (in other words: tests are asymptotically true). This approach thus may yield misleading results as illustrated below.

Consider an example (Model Scenario 1) in which we compare two corpora of the same size (N = 100,000), and the frequencies of word A:

Table 8. Model Scenario with Word A

	Corpus 1	**Corpus 2**
fq(A)	130	100
N	100,000	100,000
RelFq(A)	0.013	0.010

The ratio between the relative frequencies is 0.013/0.01 = 1.3, which means that the frequency of word A in Corpus 1 is 30% higher than that in Corpus 2. The result is significant, as the log-likelihood test statistic is 3.92, which is above the critical value (i.e., 3.84 at the 5% level).

Compare the above situation and the Model Scenario with word B in Corpus 1 and 2 below:

Table 9. Model Scenario with Word B

	Corpus 1	**Corpus 2**
fq(B)	1100	1000
N	100,000	100,000
RelFq(B)	0.011	0.010

Here, the ratio between relative frequencies is lower (1.1), which means that the frequency of word B is only 10% higher in Corpus 1 than in Corpus 2, but the log-likelihood test yields 4.76, thereby indicating that the difference is "more significant."

Ranking according to statistical significance would place word B above word A because the significance test yields a higher value for B, despite the fact that the relative difference in frequencies for word A in Corpus 1 and Corpus 2 is larger than that for word B. This leads to the conclusion that significance tests show **reliability of the difference** (given the amount of data and the observed difference) rather than its importance or prominence.[29]

Several attempts were made to overcome this issue with KW ranking. One of them is the "simple math" approach by Kilgarriff (2009), who proposes a simple ratio of the relative frequency of a word in the Ttxt to the relative frequency of the word in the RefC. This method, however, leads to another issue: what to do with the situation where a KW is not found in a RefC (henceforth RefC = 0). To avoid the problem of dividing by zero, this approach adds a constant (X) to both values as in the following formula:

(a) $$ratio = \frac{RelFq(Ttxt) + X}{RelFq(RefC) + X}$$

The value of X is important here, since different Xs lead to wide-ranging results: e.g., if X = 1, the ratio would retrieve highly obscure words; if X = 100, higher frequency words would be at the top of the list. Researcher bias is not completely removed.

Another approach to KW ranking metrics is ProcDiff, shown in (b), proposed by Gabrielatos and Marchi (2012). Their approach is based on the difference between the relative frequencies in the Ttxt and RefC:

(b) $$ProcDiff = \frac{RelFq(Ttxt) - RelFq(RefC)}{RelFq(RefC)} \times 100$$

This approach may yield misleading results when RefC = 0. If a word is absent from the RefC, the authors suggest substituting the value for RefC with an arbitrarily selected infinitely small number. True, if a word is not in the RefC, it does not necessarily mean that it is not used at all. However, by essentially treating a situation where Ref = 0 (where

[29] It must be said, however, that statistical significance is nevertheless a valuable concept in keyword analysis to identify words that merit examination. The discussion here concerns ranking of KWs, which were identified as being statistically significant.

there are no actual data to motivate prominence) as if it were a situation where Ref ≠ 0 (where there is actual data), it still runs into a problem of how to motivate the resulting numbers.

The method may even yield misleading results. Consider the following Model Scenario with words A, B, and C in a Ttxt (N = 1000); here, none of these words is present in the RefC (N = 1000000):

Table 10. Model Scenario with Words A, B, and C (RefC = 0)

Word	**Fq(Ttxt)**	**Fq(RefC)**	**ProcDiff/100**
A	30	0 (subst. by 0.01)	2,999,999
B	20	0 (subst. by 0.01)	1,999,999
C	10	0 (subst. by 0.01)	999,999

Regardless of what the substitution is, the ranking will reflect the raw frequency in the Ttxt. By doing so we conflate the data from the RefC and Ttxt, assuming that if there is no information about a word in the RefC, it is justified to substitute it with the information gained from the Ttxt. This may be especially misleading in a situation where the actual frequencies in the reference usage (i.e., the population of texts from which the RefC was sampled) for A, B, and C are not equal. For the sake of argument, consider a situation where the **actual frequencies are known (RefC ≠ 0)** and are equivalent to 0.5, 0.1, and 0.01 in the RefC.

Table 11. Model Scenario with Words A, B, and C (**RefC≠0**)

Word	**Fq(Ttxt)**	**Fq(RefC) extrapolation**	**ProcDiff/100**
A	30	0.5	59,999
B	20	0.1	199,999
C	10	0.01	999,999

This yields a completely different ranking: C is the most prominent, while A is the least prominent of those three words.

Difference Index (DIN) is a further attempt to address the issue of RefC = 0. DIN is based on the premise that it is descriptively adequate to

treat words where RefC=0 as equally prominent at the initial phase of quantitative evaluation, because the actual frequencies of these words are unknown and worthy of special attention.[30] Unlike its predecessors, this method signals to the researcher that such words require more thorough inspection.

DIN is based on Hofland and Johansson's Difference Coefficient (1982: 14).[31] The formula is similar to ProcDiff with one important enhancement in the denominator:

(c) $$DIN = 100 \times \frac{RelFq(Ttxt) - RelFq(RefC)}{RelFq(Ttxt) + RelFq(RefC)}$$

The values of DIN range from –100 to 100 with the following interpretation:

Table 12. Values of DIN

–100	The word is present only in the RefC and not in the Ttxt
0	The word occurs equally often in the Ttxt and RefC (with respect to their size)
100	The word is present only in the Ttxt

DIN was designed as a variation of Dice's coefficient, which is used for comparing sets of elements. At the core of the formula is the ratio between the difference of relative frequencies and their mean, which was extrapolated to an index ranging from –100 to 100:

(d) $$DIN = 100 \times \frac{RelFq(Ttxt) - RelFq(RefC)}{RelFq(Ttxt) + RelFq(RefC)} = 50 \times \frac{RelFq(Ttxt) - RelFq(RefC)}{(RelFq(Ttxt) + RelFq(RefC))/2}$$

DIN is immune to the problem of RefC = 0 in the denominator and yields the same values (DIN = 100) for all KWs which are present in the Ttxt only. Admittedly, this does not help researchers in deciding wheth-

[30] The absence of a word from RefC can be caused by many factors: e.g., a RefC may be too small to include some rare words, or a RefC may not be well sampled from the population, and therefore is not representative with respect to these items.

[31] There are some minor differences—Hofland and Johansson's metrics do not include the coefficient of 100. More significantly, however, they do not mention the advantage of solving the problem of zero in RefC in their 1982 study.

er the words absent in the RefC are important for the interpretation or not, but it sends a clear message that there is insufficient data, and that these KWs should be therefore treated separately with scrutiny using other types of data (e.g., qualitative discourse analysis).

Appendix 2: Top 50 KWs from the Entire Corpus of NYAs

	SYN-KWA KW word forms	DIN	TOT-KWA KW word forms	DIN
1	*přičiňme* 'let's try' (with reflexive *se*)	99.9927	*spoluobčané* 'fellow citizens, voc, nom pl'	99.8058
2	*soudružky* '(female) comrade, gen sg.; voc nom acc pl'	99.9158	*přičiňme* 'let's try' (with reflexive *se*)	99.713
3	*osvobozeneckého* 'liberating, gen sg. non-fem; acc sg masc-anim'	99.9154	*rozkvétala* 'blossomed, fem sg; neut pl'	99.713
4	*pětiletky* '5-year plan, gen sg; voc nom acc pl'	99.9101	*střízlivým* 'sober, realistic, instr sg non-fem; dat pl'	99.6557
5	*soudružské* 'comrade-like, gen, dat, loc sg. fem; voc nom pl. non-masc anim; acc pl non-neut'	99.8822	*domovům* 'homes, dat pl'	99.6176
6	*bratrskému* 'brotherly, dat sg'	99.8671	*vzkvétala* 'blossomed, fem sg; neut pl'	99.5698
7	*drazí* 'dear, voc nom pl masc-anim'	99.8624	*pozdravuji* 'I greet'	99.2841
8	*horečného* 'feverish, gen sg non-fem, acc sg masc-anim'	99.846	*vstupujeme* 'we enter'	99.2254
9	*osvobozenecký* 'liberating, voc nom sg masc; acc sg masc-inanim'	99.843	*drazí* 'dear, voc nom pl masc-anim'	98.9149

10	*pětiletce* '5-year plan, dat loc sg'	99.843	*dařila* 'succeeded (with reflexive *se*) fem sg; neut pl'	98.857
11	*spoluobčané* 'fellow citizens, voc nom pl'	99.8306	*udělejme* 'let's do'	98.7624
12	*bratrskými* 'brotherly, instr pl'	99.8068	*přeji* 'I wish'	98.734
13	*zamýšlíme* 'we think deeply (with reflexive *se*)'	99.8068	*připomeneme* 'we will remind'	98.6543
14	*rozkvétala* 'blossomed, fem sg; neut pl'	99.7947	*dopady* 'consequences, nom acc pl'	98.4789
15	*svědomitou* 'conscientious, acc instr sg fem'	99.7827	*uplynulým* 'past, instr sg non-fem; dat pl'	98.4789
16	*vědeckotechnické* 'scientific-technical, gen, dat, loc sg fem; voc nom pl masc inanim, fem; acc pl non-neut'	99.7754	xvii (number, 17)	98.4035
17	*vědeckotechnického* 'scientific-technical, gen sg non-fem; acc sg non-masc-anim'	99.7646	*novoroční* 'new year, voc nom sg; acc sg, non-masc-anim; fem obliq sg.; nom acc pl'	98.2904
18	*zdravíme* 'we greet'	99.7601	*pokročili* 'pogressed, pl masc-anim'	98.2904

	SYN-KWA KW word forms	DIN	TOT-KWA KW word forms	DIN
19	*všestranná* 'all-round, voc nom sg fem; nom acc pl neut'	99.7586	*zdravím* 'I greet'	98.2687
20	*dařila* 'succeeded (with reflexive *se*) fem sg; neut pl'	99.7568	*zamýšlíme* 'we think deeply (with reflexive *se*)'	98.2433
21	*pokrokovým* 'progressive, instr sg. non-fem; dat pl'	99.7465	*pohodu* 'comfort, acc sg'	98.1023
22	*energičtěji* 'more energetically'	99.7344	*pohodě* 'comfort, dat loc sg'	98.0083
23	*obětavé* 'dedicated, gen dat loc sg. fem; voc nom pl non-masc-anim; acc pl non-neut'	99.7304	*svědomitou* 'conscientious, acc instr sg fem'	97.9379
24	*bratrských* 'brotherly, gen loc pl'	99.7149	*posíláme* 'we send'	97.8568
25	*imperialistické* 'imperalistic, gen dat loc sg fem; voc nom pl non-masc-anim; acc pl non-neut'	99.7031	*optimismem* 'optimism, instr sg'	97.727
26	*kvalitněji* 'with higher quality'	99.7013	*poděkovat* 'to thank, inf'	97.7036
27	*střízlivým* 'sober, instr sg non-fem; dat pl'	99.6959	xvi (number, 16)	97.6428

28	xvii (number, 17)	99.6959	*vzestupný* 'mounting, voc nom sg masc; acc sg masc-inanim'	97.6334
29	*mírového* 'of peace, gen sg non-fem; acc sg masc-inanim'	99.6824	*opíráme* 'we rely on' (with reflexive *se*)	97.5867
30	*upevňování* 'stabilization non-instr sg; voc nom gen acc pl'	99.6714	*generacím* 'generation, dat pl'	97.54
31	*pozdravuji* 'I greet'	99.6652	*kvalitněji* 'with higher quality'	97.3765
32	*socialistického* 'socialist, gen sg non-fem; acc sg masc-anim'	99.6542	*přikládáme* 'we attribute to'	97.3765
33	*socialistickými* 'socialist, instr pl'	99.6525	*vážení* 'dear (lit. respected), voc nom pl anim'	97.3389
34	*bratrský* 'brotherly, voc nom sg. masc; acc sg. masc-inanim'	99.6501	*spokojený* 'satisfied, voc nom sg masc; acc sg masc-inamim'	97.3266
35	*rolníkům* 'peasant, dat pl'	99.6226	*náročně* 'with intensity'	97.3066
36	*činorodé* 'productive, gen dat loc sg. fem; voc nom pl non-neut non-masc-anim; acc pl. non-neut'	99.6019	*vašim* 'your, dat pl'	97.1668

	SYN-KWA KW word forms	DIN	TOT-KWA KW word forms	DIN
37	*vstupujeme* 'we enter'	99.5934	*zdravíme* 'we greet'	97.0273
38	*náročně* 'with intensity'	99.5929	*uplynulý* 'past, voc nom sg masc; acc sg non-masc-inanim'	96.8879
39	*přikládáme* 'we attribute to'	99.5929	*přátelé* 'friend, voc nom pl'	96.8142
40	*mírový* 'of peace, voc nom sg masc; acc sg non-masc-inanim'	99.579	*důvěrou* 'trust, instr sg'	96.7109
41	*opíráme* 'we rely'	99.5749	*nejspolehlivější* 'the most reliable, voc nom sg; acc sg. non-anim; obliq sg fem; voc nom acc pl'	96.6098
42	*socialistických* 'socialist gen loc pl'	99.5418	*opravňují* '(they) justify'	96.6098
43	*mírovému* 'of peace, dat sg non-fem'	99.5297	*prožili* 'experienced, lived through, masc-anim pl'	96.5701
44	*oceňujeme* 'we value'	99.5225	*rozvíjelo* 'developed, neut sg'	96.5543

45	*vzkvétala* 'blossomed, fem sg; neut pl'	99.5207	*tvořivá* 'creative, voc nom sg fem; voc nom acc pl neut'	96.4711
46	*tvořivou* 'creative, acc instr sg fem'	99.5153	*hodnotíme* 'we evaluate'	96.3879
47	*pokročili* 'progressed, pl masc-anim'	99.5081	*vzpomínat* 'remember, inf'	96.3325
48	*posíláme* 'we send'	99.4964	*plodem* 'fruit, instr sg'	96.2402
49	*obětavou* 'dedicated, acc instr sg fem'	99.4936	*nestraníků* 'non-party member, gen pl'	96.1941
50	*imperialism* 'imperialism, voc nom acc sg'	99.4896	*výhodnou* 'profitable, acc instr sg'	96.1941

Appendix 3: KWs with Continuous Keyness in SYN-KWA and TOT-KWA for the Entire 15-Year Period

Word forms with keyness in more than 10 years SYN-KWA	# of years with keyness/15 years	Word forms with keyness in more than 10 years TOT-KWA	# of years with keyness/15 years
naší 'our, obliq sg fem'	15	*naší*	14
lidu 'people, gen dat loc sg'	14	*našeho*	13
našeho 'our, gen sg non-fem; acc sg masc-anim'	14	*roku*	13
roku 'year, gen dat loc sg'	14	*československa*	12
socialistické 'socialist, obliq sg fem; voc nom acc sg neut; voc nom pl non-neut; acc pl non-anim-masc'	14	*lidu* 'people, gen dat loc sg'	12
národní 'national, voc nom sg; acc sg non-anim-masc; obliq sg fem; voc nom acc pl'	13	*drazí* 'dear, voc nom pl masc-anim'	11
společnosti 'society, gen, dat, loc sg; voc nom acc pl'	13	*socialistického* 'socialist, gen sg non-fem; acc sg. masc-anim'	11

československa 'Czechoslovakia, gen sg'	12	*vlasti* 'motherland, gen dat loc sg; voc nom acc pl'	11
socialistického 'socialist, gen sg non-fem; acc sg. masc-anim'	12		
světě 'world, loc sg'	12		
drazí 'dear, voc nom pl masc-anim'	11		
míru 'peace, gen dat sg'	11		
naše 'our, voc nom sg non-masc; acc sg neut; voc nom pl non-masc-anim; acc pl'	11		
našich 'our, gen loc pl'	11		
strany 'party, gen sg; nom acc pl'	11		
všech 'all, gen pl'	11		
vlasti 'motherland, gen dat loc sg; voc nom acc pl'	11		
Total	17	Total	8