JAZYKOVEDNÝ ČASOPIS, 2017, roč. 68, č. 2

GRAMMATICAL CHANGE TRENDS IN CONTEMPORARY CZECH NEWSPAPERS[1]

MICHAL KŘEN
Institute of the Czech National Corpus, Charles University, Prague, Czech Republic

KŘEN, Michal: Grammatical Change Trends in Contemporary Czech Newspapers. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 238 – 248.

Abstract: The paper presents a corpus-driven method for the detection of recent grammatical change in contemporary Czech newspapers. It is based on a large and homogeneous material (825 million tokens of a single newspaper) that covers a 23-year time span. The task is operationalised into finding the most relevant frequency change manifested by selected subsets of the Czech tagset. The results show changing proportions of parts of speech, nominal cases etc. that indicate a shift towards more "verbal" language associated with increasing informality of the newspaper register.

Keywords: modern diachrony, language change, Czech, newspaper register, corpus composition

1 INTRODUCTION

The paper aims to investigate recent grammatical change that can be observed in contemporary Czech newspapers. It presents an automatic corpus-driven method used to detect morphological features that show the most considerable diachronic shift. Finally, the results as well as the limitations of such an approach are discussed.

The paper draws on previous research done mostly on English trying to detect recent language change [1], [3], [8], [9], [12], [13], [14]. Compared to them, this study can be characterised by the following:

- it is based on large and homogeneous data;
- morphological categories (rather than the often studied individual word forms) are investigated systematically and in a corpus-driven manner; this has been operationalised into finding the most relevant frequency changes manifested by selected subsets of the currently used Czech tagset;
- evaluation of the frequency differences is carried out using Mann-Kendall test and Theil-Sen estimator.

2 DATA

It is often emphasised that research aiming to discover recent language change should be based on large and homogeneous data covered by many data points [9], [14]. This has determined selection of SYN v4 as the base corpus [11]. With its 4.3

[1] This study was written within the programme Progres Q08 *Czech National Corpus* implemented at the Faculty of Arts, Charles University.

DOI 10.1515/jazcas-2017-0033

billion tokens (3.6 billion running words), it is the largest available traditional (as opposed to web-crawled) corpus of contemporary written Czech featuring reliable metadata and large homogeneous newspaper subcorpora. SYN v4 is uniformly processed, which includes text cleanup, de-duplication and other rather technical issues, as well as lemmatisation and morphological tagging [7], [10].

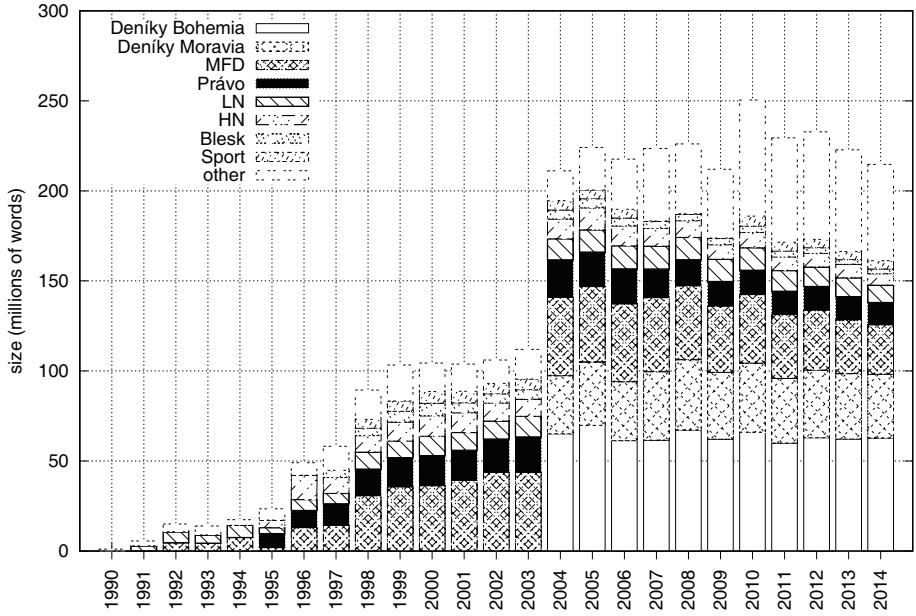

Fig. 1. Composition of the newspaper part of SYN v4

Composition of the newspaper part of SYN v4 is given in Figure 1.[2] As a rule, all the texts are incorporated in full, which means that there are always whole newspaper issues included in SYN v4. For the present study, a part of SYN v4 that contains only a single major national daily newspaper *Mladá fronta DNES* (MFD) was used. Its total size is 825 million tokens (687 million running words) and it covers the period 1992–2014, which means that it is the largest newspaper in terms of size and time span.

Subsequently, a virtual subcorpus of MFD was created for each year of the given period that was used for all the queries described in Section 3. It should be noted that some MFD issues are missing in SYN v4, especially from 1992–1995 (see Figure 2). However, this shortcoming should not distort the overall picture of the language used in MFD at the time and it is presumably outweighed by the greater number of data points available [14, p. 208].

[2] More information can be found at `https://wiki.korpus.cz/doku.php/en:cnk:syn:verze4`.

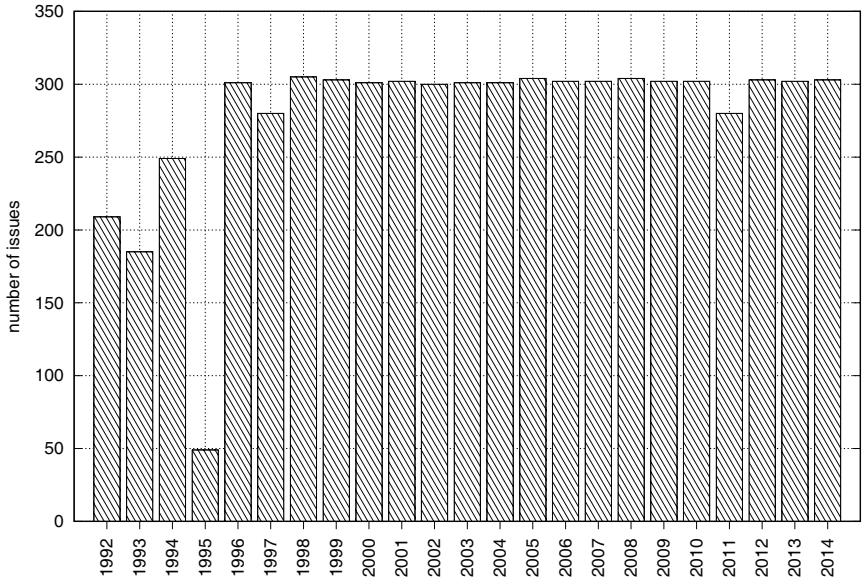

Fig. 2. Number of MFD issues per publication year in SYN v4

3 METHOD

3.1 Morphological Categories

The method is based on the Czech tagset currently used in the Czech National Corpus (CNC). The tagset draws on the original one developed by Jan Hajič [4] with some improvements and extensions.[3] Czech is a morphologically rich language and this is reflected also in the tagset: there are 4 351 different tags actually used in SYN v4. The tagset is positional, which means that each position in the tag (viewed as a string) represents a single morphosyntactic feature. For instance, one of the possible morphological tags for the word form *nejasnější* ('less clear') is `AAFS7----2N-----` which denotes the following features: adjective (A), regular adjective (A), feminine (F), singular (S), instrumental (7), comparative (2), negated form (N). For features not relevant for the given POS, '-' is used on the respective position.

The tags are very fine-grained, which means that their development over time would hardly show any convincing trend. Instead, it would be optimal to discover the trends for all possible tag combinations, and then to choose the most significant from among them. Although this procedure would guarantee that no relevant combination (in terms of the original tagset) is missed, it is also not feasible given the exponential size of the set of all possible subsets.

Therefore, instead of grouping the individual tags, *categories* were introduced that represent various morphological "dimensions", e.g. adjectives, feminine adjectives, adjectives in instrumental singular, negated adjectives, instrumental

[3] Detailed information available at `https://wiki.korpus.cz/doku.php/seznamy:tagy` (Czech only).

singular in general (without any POS restrictions), etc. The categories are defined by *regexps* (regular expressions) over the tagset that result from the expansion of variables of manually input *patterns*; the variables denote any possible value on a given position. The regexps are finally turned into the individual CQL *queries* for the Manatee query engine [15].

For instance, the pattern `${1}.*` with a variable on the first position was expanded into separate regexps for all possible values of POS, e.g. `N.*` for nouns, `A.*` for adjectives, `V.*` for verbs etc. The variables could also be freely combined to yield more complex sets of regexps for each pattern, for instance `${1}...${5}.*` with variables on the first and on the fifth position expanded into all combinations of POS and case, e.g. `N...1.*`, `N...2.*`, `N...3.*`, ... , `A...1.*`, ... , `A...7.*` etc. For some categories, CQL regular expressions were used directly (e.g. `P[PH5].*` for personal pronouns) and they were also combined with variables (e.g. `P[PH5]..${5}.*` for personal pronouns in every particular case).

The total number of input patterns was 161. After the expansion, they yielded 128 557 regexps that define the individual categories on various levels of granularity in different morphological dimensions. The regexps were turned into CQL queries simply by wrapping them into `[tag="regexp"]`. Finally, all the CQL queries were run against the individual MFD subcorpora of the SYN v4 corpus using Manatee API. For every query, a *data row* of normalised frequencies was the result, showing the development of the particular category in MFD over time. As the pattern expansion mentioned above heavily overgenerates, many resulting queries gave no results, e.g. `[tag="V...7.*"]` (verbs in instrumental). All such data rows were simply discarded and not included into the evaluation.

3.2 Evaluation

All non-zero data rows (4 628 in total) were used as an input into the statistical module that was employed to detect those with the most significant frequency development. Two methods have been used: Mann-Kendall test and Theil-Sen estimator.

Mann-Kendall is a non-parametric statistical test used to measure the correlation between ranks of two variables that is often used to assess the (upward or downward) monotonicity of the trend of the observed variable over time [6], [9]. Its values are in the <-1;1> range: 1 for perfect agreement (both variables increase/decrease simultaneously), -1 for perfect disagreement (the opposite of the above), 0 if there is no correlation observed. However, Mann-Kendall does not take into account the actual values as long as their rank order over the time remains the same. It also gives clear preference to smooth, monotonous frequency change which may not be the case in reality.

Therefore, it was supplemented by the Theil-Sen estimator that is also used for automatic detection of development trends in the Sketch Engine software (based on [5]). It is a robust linear regression method that computes a median slope of the overall trend. Theil-Sen overcomes local fluctuation in the observed trend and, at the same time, it naturally takes into account the actual frequency values (overall increase or decrease). This means that both methods complement each other.

4 Results

Evaluation of the results is complicated by the fact that most of the categories are multi-dimensional and interrelated with other ones. For instance, there is a slight increase observed for nouns in accusative, while there is a decrease for nouns in general and an increase for accusative (regardless of POS); the overall picture is presumably much more complex. At the same time, there are too many results to show as the space in this study is limited.

Therefore, two tables are presented, one for each method, with 15 items per table. The items have been selected as the most significant as detected by the given method while omitting near duplicates, e.g. `..FS2.*` vs. `..F.2.*` ('F' for feminine, '2' for genitive), where the difference is only in the (un)specification of the number ('S' for singular).

For every item, the following is given:

- exact query in terms of the CNC tagset;
- value obtained by the respective method (Mann-Kendall or Theil-Sen);
- rank according to that method;
- relative frequencies (instances per million) in 1992 and 2014 (the first and the last year of the data row);
- overall increase or decrease (trend);
- characterisation of the query (category).

Query	value	rank	i.p.m. (1992)	i.p.m. (2014)	trend	category
[tag="PDN.1.*"]	0.96	3	2604	5291	+	pronoun, demonstrative, neutral, nominative (mostly the form *to*)
[tag="PH..4.*"]	0.95	8	610	1495	+	pronoun, personal in short form, accusative
[tag="P5F.6.*"]	0.95	13	103	165	+	pronoun, personal after prep., feminine, locative (*ní*)
[tag="P[PH5].*"]	0.94	25	7605	11848	+	pronoun, personal
[tag="A...2.*"]	-0.94	32	31436	22036	-	adjective, genitive
[tag="N...2.*"]	-0.94	34	88698	75080	-	noun, genitive
[tag="PH.*"]	0.94	35	1476	3001	+	pronoun, personal in short form
[tag="AG.P.*"]	-0.93	43	937	597	-	adjective, derived from present transgressive, plural
[tag="..F.2.*"]	-0.93	46	49992	41048	-	feminine, genitive (any POS)
[tag="P[PH567].*"]	0.93	56	23302	33128	+	pronoun, personal or reflexive
[tag="P6..7.*"]	0.93	57	112	221	+	pronoun, reflexive in long form, instrumental (*sebou*)
[tag="...S2.*"]	-0.93	58	85829	69032	-	singular, genitive (any POS)
[tag="Vs.....P...P"]	-0.93	61	4477	1943	-	verb, passive participle, perfective
[tag="P5.S4.*"]	0.92	69	161	336	+	pronoun, personal after prep., singular, accusative
[tag="VB.....1F.*"]	0.92	71	319	582	+	verb, future tense, 1st person

Tab. 1. Mann-Kendall

Query	value	rank	i.p.m. (1992)	i.p.m. (2014)	trend	category
[tag="N.*"]	-1316	1	313489	285839	-	noun
[tag="....2.*"]	-1161	3	151214	128874	-	genitive (any POS)
[tag="V.*"]	1146	4	115603	142042	+	verb
[tag=".....A.*"]	1075	6	89508	116202	+	active voice
[tag="VB.*"]	912	12	49506	66430	+	verb, present or future form
[tag="V.....I"]	899	13	71988	90781	+	verb, imperfective
[tag=".....P.*"]	851	19	45479	61577	+	present tense
[tag="A.*"]	-775	23	107316	91097	-	adjective
[tag="P.*"]	727	26	60170	74222	+	pronoun
[tag="....4.*"]	699	29	117906	135418	+	accusative (any POS)
[tag="V..S.*"]	648	38	68746	84743	+	verb, singular
[tag="....1.*"]	-602	43	107619	95130	-	positive (comparison degree; adjectives and adverbs)
[tag="....3.*"]	591	44	49302	62770	+	3rd person (pronouns and verbs)
[tag="N..S2.*"]	-562	48	60579	50022	-	noun, singular, genitive
[tag="D.*"]	558	49	45980	57983	+	adverb

Tab. 2. Theil-Sen

To comment on the methods briefly, Theil-Sen tends to prefer more "global" categories (e.g. whole parts of speech), because they are more frequent. On the other hand, Mann-Kendall prefers more detailed categories (e.g. sub-part of speech combined with gender and/or case, sometimes even specific enough to single out a word form) that show monotonous development over time (by definition, monotonicity is the only evaluation criterion for Mann-Kendall).

It should be pointed out that the i.p.m. values given in both tables should be viewed as boundaries, as the i.p.m.'s for the individual years between 1992 and 2014 often develop evenly within this range (this is caused by the nature of the methods employed, especially the Mann-Kendall). Figure 3 and Figure 4 show examples of such development, one from each table, that depict one increasing and one decreasing trend. What can be observed is thus gradual, smooth and continuous frequency change of the individual grammatical categories.

In terms of the parts of speech, there is a steady increase observed for verbs, pronouns and adverbs that is complemented by the decrease of nouns and adjectives (Table 2). Even more significant change can be observed within the individual POS: Table 2 suggests that verbs in 3rd person, singular, present or future form,[4] imperfective, active voice are in the lead of the change (it is perhaps worth mentioning that all these categories constitute unmarked forms of a verb). Similarly, the increase of pronouns can be ascribed to demonstrative, personal and reflexive pronouns that often show strikingly monotonous trends (cf. Table 1 and Figure 3).

[4] Perfective verbs in Czech form future tense by their morphologically present form. This is reflected by the `VB.*` tag that denotes morphologically present verb forms.

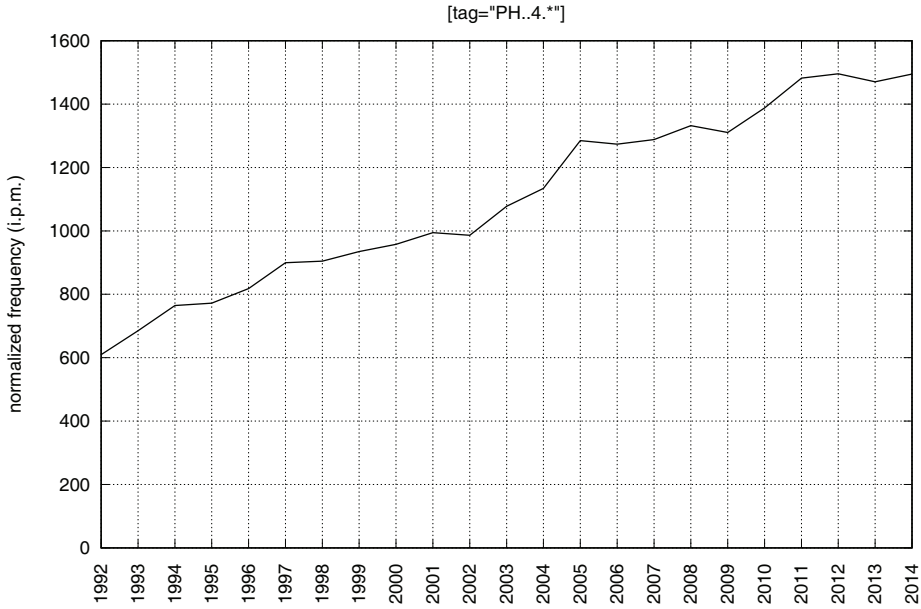

Fig. 3. Personal pronouns (short forms) in accusative

The detected trends can have various (and interrelated) causes: frequency change of the individual parts of speech suggests a shift towards more "verbal" language associated with a diversion from nominal expressions. This corresponds to the gradual move of the newspaper register towards fiction [2], [9], although one of the reasons is likely to be the growing proportion of leisure themes, interviews, weekend supplements etc. in the MFD subcorpora.

Numerous papers on recent language change also report increasing informality of the newspaper register [1], [3], [8], [9], [12], [13], [14]. This is seconded also in this study and illustrated by the decrease of rather formal expressions, namely passive participles and adjectives derived from the present transgressive (cf. Table 1).

As for the other morphological categories, there is an increase observed for accusative and decrease for genitive, regardless of POS. Given the gradual nature of this frequency shift, it is unlikely to be affected by disambiguation errors. However, one should be very cautious about its possible interpretation as an indication of a long-term typological change. Certainly more reasonable cause could simply be changing structure of MFD, perhaps also within its individual sub-registers, which will be discussed in the following paragraphs.

In order to investigate the composition of MFD and its possible influence on the presented results, the newspaper sections markup newly introduced in the SYN-series corpora was used. The information about the sections is available for all articles published in major newspapers (including MFD) since 2010. It is based on the original section titles taken over from the publishers that have subsequently been

unified and classified into the sections available as a part of the newspaper article metadata (with a small percentage of section titles remaining as unclassified).

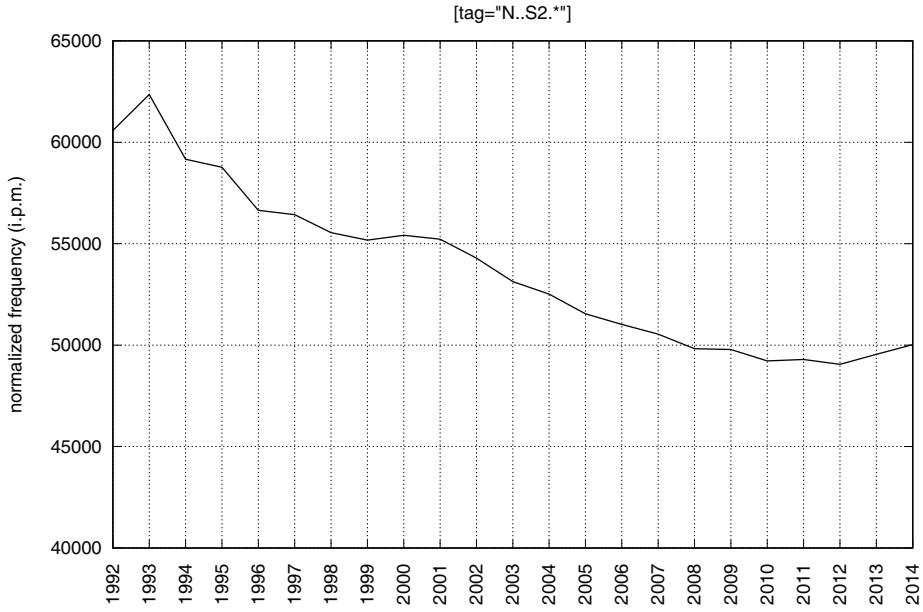

Fig. 4. Nouns in genitive singular

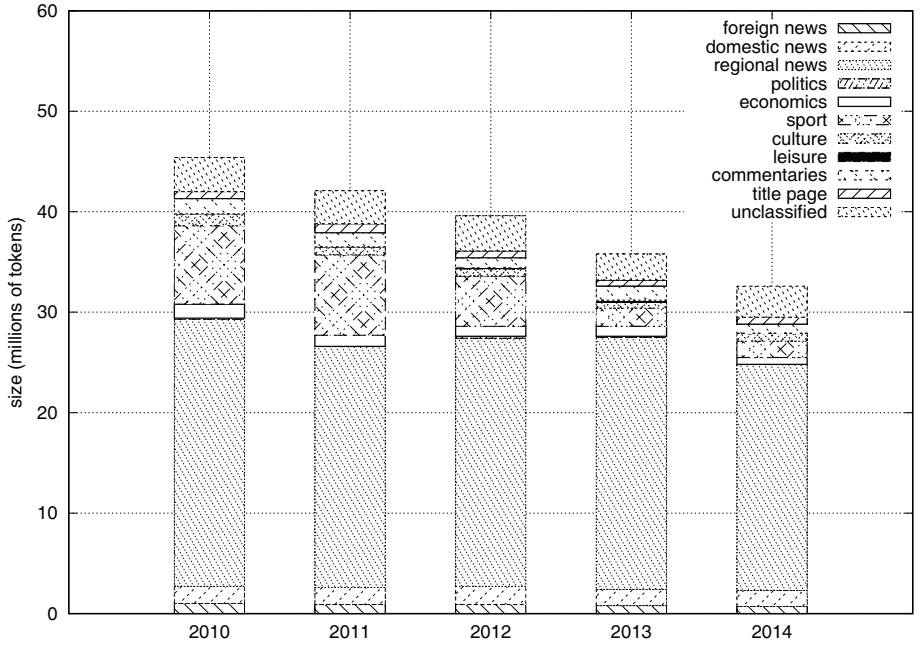

Fig. 5. Newspaper sections in MFD

Composition of MFD in terms of the individual sections in 2010–2014 is shown in Figure 5. There are two immediate observations to be made: first, the most prominent part of MFD are regional news, and second, the average size of an MFD issue is decreasing (as there is about 300 issues per year, see Figure 2).

The decreasing size of MFD in SYN v4 is the result of the decreasing volume of MFD texts received from the publisher's archive and available at the beginning of the corpus processing pipeline. As for the prevalence of regional news, the general policy (already mentioned above) is to include full texts or newspaper issues into SYN v4. However, MFD is sold in numerous regional versions that differ mostly in their regional news section. All the regional versions of MFD from the same day make up one issue on the input, which is then subject to de-duplication procedures on article level before its inclusion into SYN v4 [7, p. 160]. This affects mainly the non-regional articles typically removed as identical across the regional versions, while the bulk of the regional news remain as the prevailing part of MFD in SYN v4.

Although, there are no data on newspaper composition available for periods before 2010, it can be concluded that the results of this study are presumably not caused by the changing structure of MFD, but rather by gradual shifts within its individual sub-registers, most notably the regional news.

5 SUMMARY

The paper presented a corpus-driven method for detection of recent change in morphological categories that can be observed in contemporary Czech newspapers. The trends can be characterised as a shift towards more "verbal" language associated with increasing informality of the newspaper register.

The study certainly has its limitations. First, only categories that resulted from manually selected patterns have been considered in the evaluation. This means that some of them may have been left out, either by unintentional omission, or simply because the given combination was not considered potentially relevant.

Another limitation pertains to the morphological tagging that underlies the individual categories and that may not be ascribed only to the disambiguation accuracy. For instance, there has been incidentally discovered only a slight decrease of nominal (short) forms of adjectives (`AC.*`) during the examination of the results. This is in contradiction with their gradual replacement by their long counterparts in contemporary Czech. A key to the explanation is the word form *rád* ('glad'): from diachronic point of view, it is a nominal form of an adjective and it is also tagged as such. However, it is fossilised and contemporary Czech descriptions often treat it as an adverb. Since it is both very frequent and typical of informal language and topics, its increase almost compensates for the (quite significant) decrease of all other nominal forms. Figure 6 shows the resulting plot as a confluence of the two factors.

Last but not least, the study aimed at the detection of grammatical change trends in Czech newspapers. However, it analysed only one national daily newspaper (MFD) and came to the conclusion that the analysis is for the most part based on the regional news within MFD. This should not be seen as a shortcoming, as studies

based on large and homogeneous data from restricted domain certainly have their value, although the results may not be as general as one would wish: "*The advantage of working with single source data ... is that, although the claims that can be made are necessarily limited, they are securely grounded*" [14, p. 216]. At the same time, the study has confirmed that the major challenge for research on recent language change is the data. CNC thus aims to continuously build corpora also from other domains to provide the research community with constantly growing material that could eventually bring corpus-derived insights into the nature of language change.

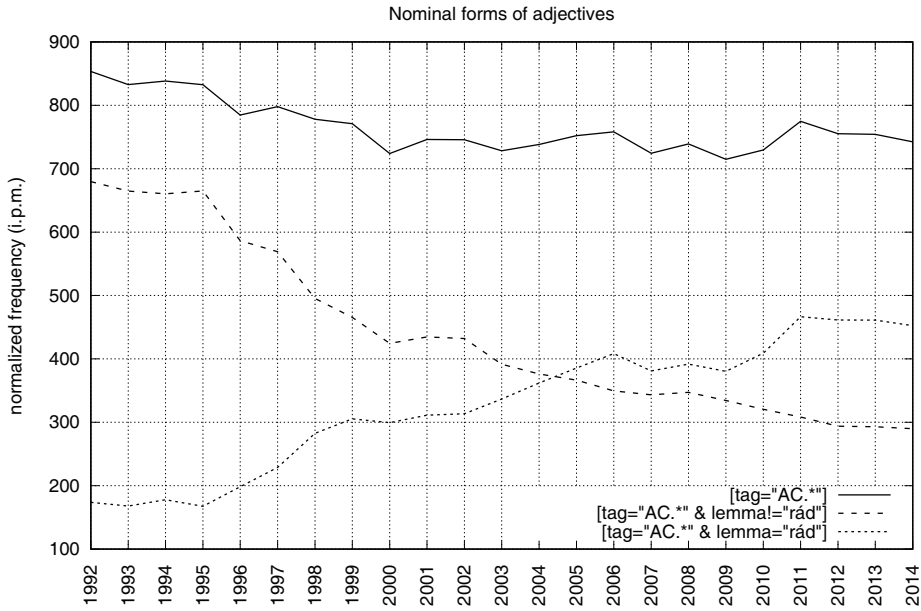

Fig. 6. *rád* vs. all other nominal forms of adjectives

References

[1] Baker, P. (2009). The BE06 Corpus of British English and recent language change. *International Journal of Corpus Linguistics*, 14(3):312–337.

[2] Bartoň, T., Cvrček, V., Čermák, F., Jelínek, T., and Petkevič, V. (2009). *Statistiky češtiny*. Nakladatelství Lidové noviny, Praha.

[3] Duguid, A. (2010). Newspaper discourse informalisation: a diachronic comparison from keywords. *Corpora*, 5(2):109–138.

[4] Hajič, J. (2004). *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Karolinum, Praha.

[5] Herman, O. and Kovář, V. (2013). Methods for Detection of Word Usage over Time. In *Seventh Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2013*, pages 79–85, Tribun EU, Brno, Czech Republic.

[6] Hilpert, M. and Gries, S. T. (2009). Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing*, 24(4):385–401.

[7] Hnátková, M. et al. (2014). The SYN-series corpora of written Czech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 160–164, ELRA, Reykjavík, Iceland.
[8] Hundt, M. and Mair, C. (1999). 'Agile' and 'Uptight' Genres: The Corpus-based Approach to Language Change in Progress. *International Journal of Corpus Linguistics*, 4(2):221–242.
[9] Křen, M. (2013). *Odraz jazykových změn v synchronních korpusech*. Nakladatelství Lidové noviny, Praha.
[10] Křen, M. et al. (2016). SYN2015: Representative Corpus of Contemporary Written Czech. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC 2016*, pages 2522–2528, ELRA, Portorož, Slovenia.
[11] Křen, M. et al. (2016). SYN corpus, version 4 from 16. 9. 2016. Ústav Českého národního korpusu FF UK, Praha. Accessible at: `http://www.korpus.cz`.
[12] Leech, G. (2004). Recent grammatical change in English: Data, description, theory. In *Advances in Corpus Linguistics. Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23)*, pages 61–81, Rodopi, Amsterdam, Netherlands.
[13] Mair, C., Hundt, M., Leech, G., and Smith, N. (2002). Short term diachronic shifts in part-of-speech frequencies: A comparison of the tagged LOB and F-LOB corpora. *International Journal of Corpus Linguistics*, 7(2):245–264.
[14] Millar, N. (2009). "Modal verbs in TIME: Frequency changes 1923–2006". *International Journal of Corpus Linguistics*, 14(2):191–220.
[15] Rychlý, P. (2007). Manatee/Bonito – A Modular Corpus Manager. In *First Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2007*, pages 65–70, Brno, Czech Republic.