Five Years of Language Services

Zuzana Nevěřilová

Natural Language Processing Centre
Faculty of Informatics
Botanická 68a, Brno, Czech Republic

Abstract. The paper analyzes the usage and patterns observed within Language Services, an aggregation website offering various APIs for natural language processing tasks. Over five years, logs for eight services were collected, allowing for a detailed investigation into the utilization of individual services. Overall, the APIs were used nearly 80 thousand times. The paper focuses on tracking service-specific usage, identifying common trends, detecting potential misuse, and examining error occurrences. The findings provide insights for possible service improvements and future enhancements.

Keywords: declension, tagging, topics, API, log analysis

1 Introduction

Language Services is an aggregation website that provides APIs[1] for various NLP tasks. It was launched in April 2018 without any wide publicity. This paper presents observations from the API logs to see how the services were used. We focused on the number of requests, the number of requests per IP address, and how appropriate the use was. We also discovered that some services did not work for particular inputs or periods.

Section 2 provides an overview of the website usage, and Section 3 describes individual services and observations about their use. Section 4 shows what IP addresses called the services. It can be seen that many users used only one or two services, while others tried all services with a small number of requests. Section 5 summarizes observations of individual services and proposes further improvements.

2 Language Services

The Language Services[2] provides 13 APIs for Czech and 2 APIs for English. These are:

1. `majka` – morphological analysis, 2. `logic` – intensional logical analysis, 3. `diacritics` – diacritic restoration, 4. `inflection` – inflection of words,

[1] Application Programming Interface

[2] `https://nlp.fi.muni.cz/languageservices/`

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2023, pp. 11–22, 2023. © Tribun EU 2023

5. topics – find topics in text, 6. phrases – extraction of (sub)phrases, 7. polite – detection of rude words, 8. vocative – generator of vocative forms, 9. sholva – shallow ontology for Czech words, 10. gen – word forms generator, 11. get location – find location names in text, 12. declension – declension of noun phrases, 13. tagger – tagging of Czech and English, 14. hello – Example service

Unfortunately, we do not have log files for all services; however, for the majority, we do. We investigate eight logs of "real" services (we omit the Hello service); the tagger service is investigated in 3.4 for both languages.

3 Usage Statistics by Service

The Language Services were launched in March 2018 with declension, tagger, polite, diacritics, vocative, get location, topics. The majka service was added in June 2018, gen, logic, and phrases services were added in Fall 2019. We collected usage statistics for the eight services with logs as shown in Table 1. The unknown service means users requested a non-existing service.

Table 1: Language Services usage statistics

service name	requests
declension	10,676
diacritics	12,312
get location	1,034
hello	422
majka	5,673
polite	1,649
tagger	38,759
topics	7,607
unknown	25
vocative	1,071
total	79,228

There were periods when the services were not fully functioning. In Table 2, we provide an overview of error types and the error's last occurrence. It seems that some errors were fixed meanwhile. Since we log the inputs, the error logs will serve for debugging the services.

3.1 The diacritics service

The service restores diacritics in Czech texts. Since there is massive ambiguity in words without diacritics (e.g., "muzu" might be a form of "můžu" (*I can*), "múzu" (*source of inspiration*), or "mužů" (*man*)), the method is based on n-gram frequencies in the corpus [2]. The diacritic restoration is much more accurate within a context that can reduce the ambiguity.

Table 2: Number of type of errors in the requests

service name	# of errors	error types	last occurrence
declension	3	UnicodeEncodeError UnicodeDecodeError	2022-05-08
diacritics	10	UnicodeEncodeError AttributeError	2023-05-24
polite	11	sre constants (regex error)	2020-03-22
tagger	8	NameError	2023-05-25
topics	681	NameError, TypeError	2021-03-06
vocative	2	IndexError	2023-09-22

The service was used 12,312 times; after filtering out the example requests, the number dropped to 11,869. Next, we filtered out 16 requests with invalid input (missing the `text` parameter). The number of unique requests was 10,292. The service usage is shown in Figure 1

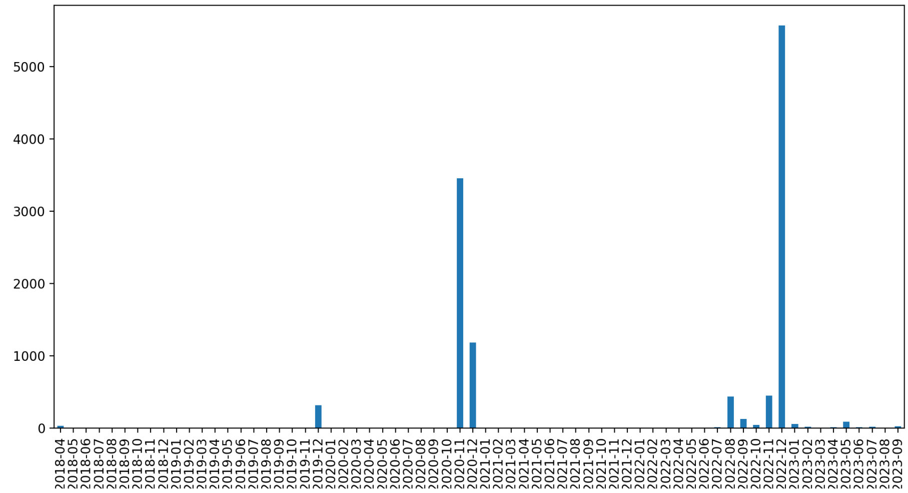

Fig. 1: Usage of the `diacritics` service (requests without example requests)

In November and December 2020, the service was called 3,236 times by Slovak IP addresses (`orange.sk` and `soitron.sk`). The service was called with single Czech words (e.g., "predlozim" (*I will present*), "jeste" (*still*), "strelec" (*sniper*)).

The highest peak occurred in December 2022, when various IP addresses at `amazonaws.com` used the service more than 5,000 times. It is unclear why the service was used since the inputs contained single Czech words with (!) diacritics.

Overall, the service was mainly misused. Using the service for single words leads to less accurate results, so we should consider providing a disclaimer.

3.2 The declension service

The service for declension provides word form for single nouns or noun phrases. The input is the text and its input case, desired output case, and output number (the same if not provided). The function of the method is described in detail in [1].

The declension service was called 10,676 times; however, 4,947 requests exceeded the daily limit. After filtering out the example requests, there were 6,142 requests. We also filtered out 99 invalid requests, such as missing parameters. The service logs contained 2,655 unique requests.

In 2018-04, the service was tested (480 requests) and used by an IP address at ncr.com (275 times), apparently for testing purposes. The service usage is in Figure 2.

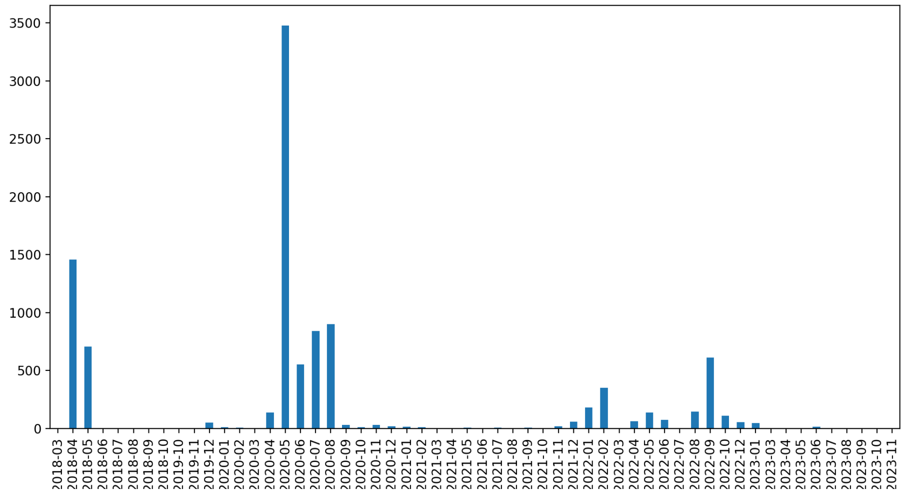

Fig. 2: Usage of the declension service

In May and June 2020, the service was called by IP addresses from Slovakia slavconet.sk and t-com.sk (666 times). In July and August 2020, the service was used by IP addresses from shawcable.net (448 requests), rogers.com (557 requests), and tmcz.cz (333 requests). The requests from tmcz.cz contained noun phrases apparently from newspapers converted to locative case (e.g., "mezinárodní filmový festival" (*International Film Festival*), "britská královská rodina" (*British royal family*), "ministr zahraničí Tomáš Petříček" (*Tomáš Petříček, Minister of Foreign Affairs*), "zmizelý Kim Čong-un" (*the missing Kim Jong Un*)). The requests from rogers.com contained complete declension of single words, primarily numerals (e.g., all forms of the word "patnáctý" (*fifteenth*)). The requests from the slavconet.sk IP addresses were about the locative case of Czech town and village names (sorted alphabetically). The requests from `t-com.sk` aimed to obtain genitive cases of job positions (also sorted alphabetically). The inputs were single words such as "učitel" (*teacher*), "badatelka" (*researcher*), "basista" (*bass player*), "dřevostavitelka" (*woodworker*), "džihadista" (*jihadist*), many of them in their male and female forms (similar to English word pair actor/actress).

A smaller peak appeared in February 2022, when the service was used 212 times by a `ssakhk.cz` IP address to obtain genitive cases of month names (such as "cerven" – *June* without diacritics). The same IP address used the service 515 times between November 2021 and June 2022.

Another peak in the graph was in 2022-09 when the service was used 503 times by a `vodafone.cz` IP address. The service seems to be used for obtaining the base form (nominative) of various phrases from an encyclopedia.

It seems that the service is well understood and not misused. However, for some inputs, the outputs were not correct. These inputs will serve for further improvements of the service.

3.3 The `majka` service

The Czech morphological analyzer `majka` [6] is widely used, including its Python binding[3]. The analyzer provides several modes of operation; the default option is to return tags and lemmata for given input words. Only this option is provided via the Language Services. Also, the original analyzer has several morphological dictionaries[4], but only the Czech dictionary is provided via the API.

The service was used 5,673 times; after filtering out the example calls, there were 4,893 requests, 3,918 being unique (10 invalid inputs were filtered out). The service usage is in Figure 3.

We investigated the use of the `majka` service and realized the API is used in a Czechitas Digital Academy project[5]. The project launch and testing are reflected in 2019-12 peak when the service was requested 3,047 times from IP addresses from `amazonaws.com`, meaning the project was implemented in the AWS Cloud. We did not further investigate the project code. However, we noticed the service was used for individual words in short periods (seconds), even though it can be applied to a list of words.

In April 2020, the service was called 388 times from the IP address `ujezd.net`. Surprisingly, the service was used to obtain grammar tags for words from the political agenda of the ANO political party.

The highest peak was in 2022-12 when various IP addresses called the service more than 3,000 times at `amazonaws.com`. The inputs were diverse single Czech words, some being standard (e.g., "diamanty" (*diamonds*), "krabici"

[3] `http://pypi.org/project/majka/`

[4] See `https://nlp.fi.muni.cz/ma/`

[5] The project *Kategorizace firem podle klíčových slov* from Fall 2019 available at `https://rukkait.blogspot.com/2019/12/v-behaviorurldefaultvmlo.html`.

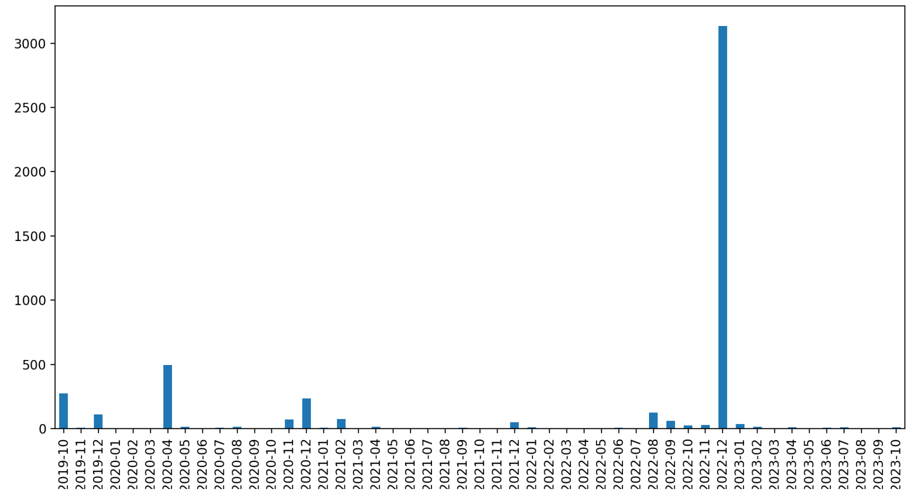

Fig. 3: Usage of the `majka` service (requests without example requests)

(*box*)), others being non-standard ("sneholak", probably *snowman*, "kávama", probably *coffees*).

We checked that most of the time, the service was used for single words (4,336 requests), 145 times, it was used for 2–9 words, 288 times, it was used for 10–100 words, 114 requests containing more than 100 words (note the input limit is 1000 characters). For future improvements, we will inform the users of the possibility of simultaneously processing more than one word.

3.4 The `tagger` Service

The tagger for Czech is based on the *desamb* tagger [5]. For English, the service uses the TreeTagger implementation [3]. Although TreeTagger supports English, French, German, and Italian, only the English version is implemented in the API service.

The tagger was used 38,759 times; after filtering out example requests, there were 38,083 requests. We removed 472 errors (input errors such as missing parameters and output errors such as no vertical was output by the tagger). Finally, from the 37,611 requests, there were 27,423 unique requests. The tagger was used 153 times for English (with the `lang=en` parameter), and the rest was for Czech. Apart from small samples of English texts (sentences such as "How are you?"), there was one request on January 9–10, 2020, that in 29 requests sent twice the text of Goosey Gazette[6]. The service usage is in Figure 4.

The tagger was used 3,125 times in May 2020 by a Slovak IP address `t-com.sk` for recognizing grammar tags in Czech job position names. The job was run in parallel with the exact requests as the `declension` service from May 13–16, 2020.

[6] `https://community.failbettergames.com/t/the-goosey-gazette/18986`

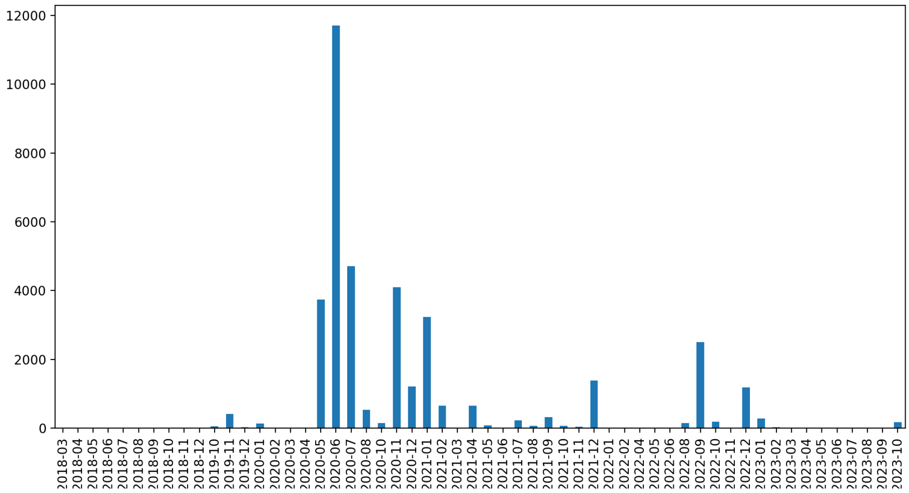

Fig. 4: Usage of the `tagger` service (requests without example requests)

From May to July 2020, various IP addresses called the service more than 16,200 times from `o2.cz`. The input texts were of a reasonable length (from single sentences to paragraphs), with topics from newspapers, the Bible, and probably fiction.

The 2020-11 peak is caused by 3,247 requests by an `orange.sk` IP address, calling only one-word requests. A similar request was made in December 2020 by another Slovak IP address from `soitron.sk`, sending single-word requests with Czech words sorted alphabetically. The author of this paper uses the tagger service in her teaching, so we are aware of the 556 requests containing parts of the Czech poem *Máj* used in the teaching. The January 2021, December 2021, and December 2022 contain mostly tagging of the poem *Máj*.

The 2022-09 peak is caused by more than 1,200 requests by IP addresses from `o2.cz` that sent texts about literature, history, and politics.

3.5 The `topics` Service

The service performs a partial syntactic analysis to discover noun phrases. Next, it converts the noun phrases to nominative case (using the underlying application of the `declension` service). Finally, it scores the noun phrases (by frequency and occurrence of proper nouns).

The service was used 7,607 times. After filtering out the example requests, there were 7,104 requests; after removing 19 invalid requests (missing `text` parameter) and duplicates, there were 4,981 unique requests. The service usage is shown in Figure 5.

In January and February 2020, the service was used by IP addresses at `tmcz.cz`. The input texts were often too short to output at least one topic. The

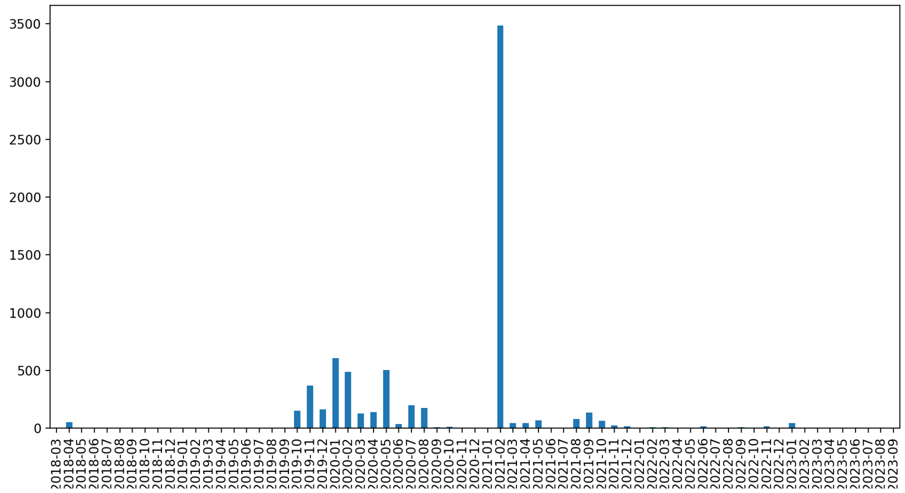

Fig. 5: Usage of the `topics` service (requests without example requests)

most common topics were "centrum" – *center*, "restaurace" – *restaurant*, "život" – *life*, "kvíz" – *quiz*, "pěkná výstava" – *nice exhibition*.

In 2021-02, the service was requested 3,190 times by a `vodafone.cz` IP address. Almost all the requests contained short texts concerning the "political restart for Czechia" – a concept by Mikuláš Minář (mentioned in the texts) and his political movement "Milion chvilek pro demokracii" (*Million moments for democracy*). The most frequent topics in this set of requests were "politika" (*politics*), "změna" (*change*), "lidé" (*the people*), "svoboda" (*liberty*), "pravda" (*truth*), "demokracie" (*democracy*), most frequent multi-word expressions were "naše země" (*our country*), "noví lidé" (*new people*), "slušní lidé" (*decent people*), "Česká republika" (*the Czech Republic*), "slušná politika" (*decent politics*), "naše děti" (*our children*), "změna politiky" (*change of the politics*), "lepší budoucnost" (*a better future*).

3.6 The `get location` Service

To discover the location mentioned in the input text, the service uses a named entity recognition (NER) implementation for Czech [4].

The service was used 1,034 times; 569 requests differed from the example request. After filtering out one invalid request (without the `text` parameter), there were 352 unique calls. Most of the calls from 2018-04 were for testing purposes.

Apart from the initial testing, the service was not used, so we do not provide a figure. After 2020, the service was used only 38 times. Some of the input texts were quite long (e.g., "Ahoj kde mohu zaparkovat právě stojí na Čápkova 43" – *Hello, where can I park a car near the Čápkova 43*).

The low usage of the service suggests stopping offering it or generalizing it by publishing a state-of-the-art NER service.

3.7 The `polite` Service

This service is based on a simple list of regular expressions describing rude Czech words.

The service was used 1,649 times. When we excluded the example use, the service was used 1,220 times, from which 613 requests were unique. The service usage can be seen in Figure 6.

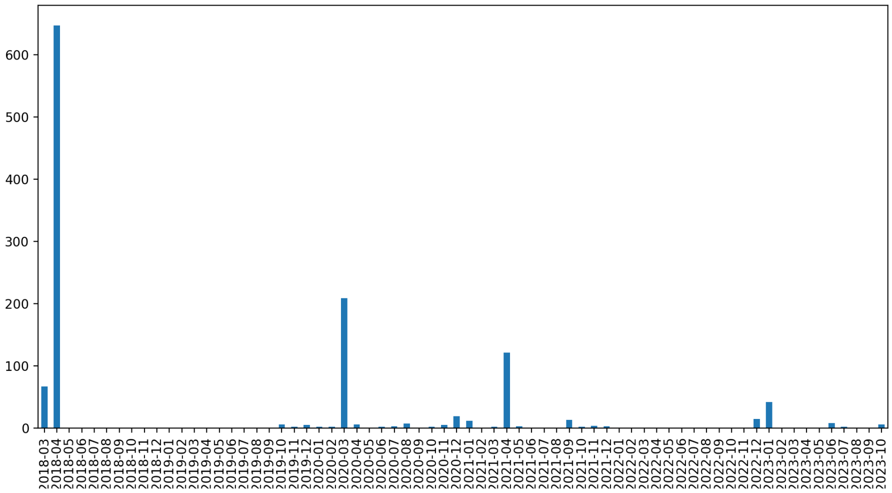

Fig. 6: Usage of the `polite` service (requests without example requests)

Apart from big testing at the service's launch, it was used 206 times in 2020-03 by an address at `vodafone.cz`. another peak is in 2021-04, where the service was used 121 times by someone at `amazonaws.com`. Both peak usages were apparently a filter, where, in fact, very few rude words appeared (only "blbec" – *dumb* and its derivatives).

3.8 The `vocative` Service

The `vocative` service generates vocative forms for Czech person names. In fact, it is a subset of the `declension` service. In contrast to the rest of the declension procedures that are based on the `majka` morphological analyzer, the declension of proper nouns is based on separate dictionaries. A similar service exists[7]. The service usage is in Figure 7.

[7] `https://sklonuj.cz/generator-osloveni/`

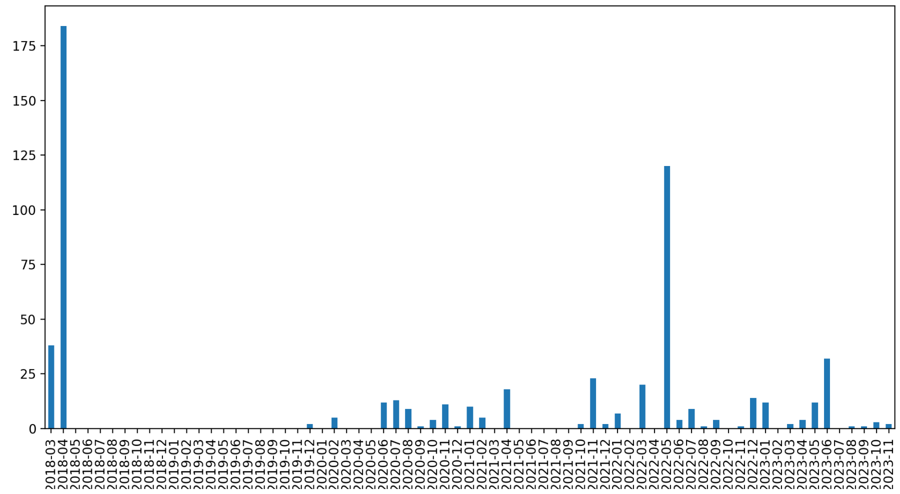

Fig. 7: Usage of the `vocative` service (requests without example requests)

The service was requested 1,071 times, 589 times with other than example input. There were 2 errors and 187 unique requests. Most of the usage was real person names converted correctly into vocative. Example requests are:

- Dagmar – incorrect vocative Dagmare
- Ivo Václav Hawiger – incorrect vocative Ive Václave Hawigere
- Klaus Mueller – incorrect vocative Klae Muellere
- Pepek Námořník – correct vocative Pepku Námořníku
- Jindřich Mořeplavec – incorrect vocative Jindřichu Mořeplaveci
- Jana Malá – correct vocative Jano Malá
- Martínek – correct vocative Martínku
- milada horáková – correct vocative milado horáková
- Admin – correct vocative Admine
- Aleksandra Pavlovna Vysockaja – correct vocative Aleksandro Pavlovno Vysockaja

The usage peaks were at the beginning when the service was launched and tested by 6 different IP addresses, then in 2022-05 when the service was used 117 times by an IP address at `selfnet.cz`.

4 Statistics of IP Addresses

During the observed period, 2,164 different IP addresses used the Language Services. In this Section, we investigated the origins of the requests. Although it is not possible to find real people or organizations from most of the IP addresses, we grouped the IP addresses based on our knowledge. The `amazonaws.com` requests are requests from programs stored in the Amazon AWS Cloud, meaning someone is using the APIs inside their own programs. A similar situation is with the `googleusercontent.com` IP addresses; these are API calls from Google Colabs used in teaching.

Table 3: Number of requests per IP address

IP address	declension	topics	tagger	majka	get location	polite	diacritics	vocative
amazonaws.com				634			1016	
amazonaws.com			1312			242		
amazonaws.com				642			1002	
amazonaws.com							1438	
amazonaws.com				1084			892	
amazonaws.com				648			994	
amazonaws.com				1296			2024	
amazonaws.com				602			978	
cvut.cz		1502	2					
ssakhk.cz	1030							
soitron.sk			2136	22			1582	
orange.sk			6496				6474	
t-com.sk	6896		6250					
o2.cz			2446					
o2.cz			1394					
o2.cz			15648					
o2.cz			6692					
o2.cz			1352					
o2.cz			1354					
tmcz.cz	2	1004				6		
tmcz.cz	970	712	252	18	6	6	4	4
vodafone.cz	12	82			860	1178	2	326
vodafone.cz	2	6506	8	2				
vodafone.cz	1228		1346					
vodafone.cz	1384						2	
ncr.com	1018						60	
shawcable.net	1260							
rogers.com	1544			2				

To our knowledge, Language Services are used in student projects at Masaryk University. Moreover, the `majka` service is used in the Czechitas Digital Academy project (see Section 3.3). Other schools used the Language Services as well (`cvut.cz` and `ssakhk.cz` is a university and a high school, respectively).

Other IP addresses are owned by Internet providers in Czechia, Slovakia, and worldwide. We cannot conclude anything. Table 3 shows domains of IP addresses that requested Language Services more than 1000 times. It can be clearly seen that `tagger` is the most popular service for Slovak IP addresses, `majka` and `diacritics` are most widely used in other applications. The IP addresses at `o2` and vodafone might belong to the same subject as both companies use a dynamic IP address assignment.

On the other hand, among the 135 IP addresses that requested more than six different services, only five sent more than 200 requests in total. This indicates another typical behavior – someone uses the APIs for experiments but not for further benefit.

5 Conclusion and Future Work

After five years of running the Language Services, it can be seen the service is known. Since it started as a toy project, there are no standard functions such as the health check or various HTTP codes for various events (e.g., 429 Too many requests). Instead, the service returns HTTP status 200 (OK) and the message inside the response. We plan to improve the service on this technical level.

For individual services, we collected enough data about how they are used. It seems reasonable to explain further what the service is good for to avoid inefficient use. In the future, we will focus on more comprehensible documentation and improvement of individual services (declension, evocative, tagger). Also, other services should store the log information for future usage analysis.

Acknowledgments. This work has been partly supported by the Ministry of Education of CR within the LINDAT-CLARIAH-CZ project LM2023062.

References

1. Nevěřilová, Z.: Declension of Czech Noun Phrases. In: Radimský, J. (ed.) Actes du 31e Colloque International sur le Lexique et la Grammaire. pp. 134–138. Université de Bohême du Sud à České Budějovice (République tchèque), České Budějovice (2012)
2. Rychlý, P.: Czaccent - simple tool for restoring accents in Czech texts. In: Aleš Horák, P.R. (ed.) 6th Workshop on Recent Advances in Slavonic Natural Language Processing. pp. 15–22. Tribun EU, Brno (2012), https://nlp.fi.muni.cz/raslan/2012/paper14.pdf
3. Schmid, H.: TreeTagger (2014), http://hdl.handle.net/11372/LRT-323, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University
4. Sedlář, L.: Hybridní systém pro detekci pojmenovaných entit v českém textu. Diploma thesis, Masaryk University, Faculty of Informatics (2014), https://is.muni.cz/th/rij93/
5. Šmerk, P.: Unsupervised learning of rules for morphological disambiguation. In: Sojka, P., Kopeček, I., Pala, K. (eds.) Text, Speech and Dialogue. pp. 211–216. Springer Berlin Heidelberg, Berlin, Heidelberg (2004)
6. Šmerk, P.: Fast morphological analysis of Czech. In: Proceedings of the Raslan Workshop 2009. Masarykova univerzita, Brno (2009), https://nlp.fi.muni.cz/raslan/2009/papers/13.pdf