Lemmatization of Czech and Croatian Noun Clusters for Terminology Extraction

Marek Blahuš[1]
and Katarína Petreková[1,2]

[1] Lexical Computing, Brno, Czech Republic
[2] Faculty of Arts, Masaryk University, Brno, Czech Republic
firstname.lastname@sketchengine.eu

Abstract. During terminology extraction, terms discovered in corpora are presented in their canonical form. Lemmatization of multi-word terms consisting of noun clusters can be ambiguous due to the lack of information on their internal structure. In this paper, we show that grammatical case alone is often not sufficient for the construction of canonical forms of noun clusters. We focus on two-noun clusters in the genitive, which are the most frequent type with ambiguous parsing. Based on corpus research, we design rules that make use of multiple morphological categories to improve the lemmatization of noun clusters found in Czech and Croatian corpora. In addition to case, we also take note of gender, animacy, and whether the noun is a proper noun. The improvements lead to more accurate and more unified forms of the terms produced during terminology extraction for these two languages in Sketch Engine.

Keywords: noun clusters; lemmatization; terminology extraction

1 Introduction

1.1 Terminology Extraction

Terminology extraction is an important feature of NLP toolkits, such as the Sketch Engine corpus manager [7], to the extent that a dedicated web interface called OneClick Terms [1] has been developed to showcase this functionality while hiding all the background complexity of corpus building, text alignment and the actual term extraction.

Sketch Engine achieves high-quality results by combining a general-purpose scoring algorithm [5] with language-specific terminology extraction grammars [3]. A terminology extraction grammar (or *term grammar*) is a set of rules which define the lexical structures that should be included in term extraction. An example of such a rule is shown in Figure 1.

1.2 Noun Phrases and Noun Clusters

Terms are mostly formed by noun phrases. Previous studies [8][4] provide detailed analyses of Slavic noun phrases. Fusional languages often, but not

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2025, pp. 165–172, 2025.

```
define(`noun', `[tag="k1.*"]')
define(`noun_genitive', `[tag="k1.*c2.*"]')
define(`adj_genitive', `[tag="k2.*c2.*"]')
define(`agree_number', `$1.n=$2.n')

*COLLOC "%(1.lemma) %(2.word) %(3.word)"
1:noun 2:adj_genitive 3:noun_genitive & agree_number(2,3)
# example: redukce červených očí
```

Fig. 1: Simplified example of a term grammar rule for Czech, along with the definitions of the used macros (at the top). The head noun in position 1 is output as the lemma, the adjective and noun in positions 2 and 3 must agree in number. The *ajka* tagset is used, cf. Table 1. The indicated example term translates as "reduction of red eyes".

always, prefer noun clusters to prepositional phrases, as inflection allows them to express relationships using cases. Therefore, noun clusters, i.e. noun phrases composed of several nouns, appear among terms. The relationships between the individual nouns that form a noun cluster influence how the noun cluster should be lemmatized, i.e. how its canonical form should be derived.

The expected result of terminology extraction are multi-word terms listed in their canonical forms, which are often different from the inflected forms in which these terms occur throughout the corpus. However, practice has showed that the determination of canonical forms for noun clusters is not straightforward. In this paper, we show that grammatical case is not the only factor in the construction of canonical forms of noun clusters, and that also other morphological categories of the participating nouns must be taken into account during lemmatization.

1.3 Lemmatization

When constructing the canonical form of a noun phrase, the head of the noun phrase always gets lemmatized. Constituents of the noun phrase which agree with the head also get turned into lemmas. Other constituents are left unchanged.

In text corpora, noun phrases do not always occur in the canonical form, and common corpora do not typically reveal the full internal structure of a noun phrase, because only shallow parsing is usually performed, and the exact dependencies between the constituents of a noun phrase may therefore remain ambiguous [3].

Specifically, in noun clusters, agreement cannot reliably be detected merely by the nouns being in the same case. Consider the following two possible lemmatizations of the Czech noun cluster *malíře Muchy* (painter-GEN Mucha-GEN):

1) *malíř* *Mucha*
painter-NOM Mucha-NOM
'painter Mucha'

2) *malíř* *Muchy*
painter-NOM Mucha-GEN
'painter of Mucha'

In (1), both nouns get lemmatized (the canonical form is of type *lemma* + *lemma*), because it is believed there exists an agreement between them. The satellite (the second word) is thought to be a noun adjunct of the head (i.e. *Mucha* is the name of the painter).

In (2), only the first noun (the head) gets lemmatized, while the second noun is left as is (the canonical form is of type *lemma* + *word*). The satellite is thought to be a complement, i.e. a non-agreeing category, in this case in the genitive (*Muchy* expresses who is depicted in the painter's paintings, or whose style the painter imitates). The fact that the two nouns share the same case in this particular occurrence is deemed only coincidental.

Indeed, **noun clusters used in the genitive** seem to be the most difficult to lemmatize accurately. There are two possible ways to parse them, and therefore two ways to lemmatize them; and, in some cases, like the above, both lemmatizations are theoretically possible. However, corpus search reveals that the canonical form type *lemma* + *lemma* is much more appropriate in this case. In other noun clusters, such as *malíře pokojů* (painter-GEN.ANIM.SG room-GEN.INANIM.PL), only one plausible lemmatization exists (the *lemma* + *word* type in this case, i.e. 'painter of rooms', 'house painter').

While it can be observed by a corpus search that the prevailing canonical form type for a two-noun cluster is *lemma* + *word*, more accurate results can be obtained if we take into account all the information on the nouns that is available in the corpus.

1.4 Available Information

We studied two-noun clusters and their appropriate lemmatization in the following corpora:

- for Czech, the Czech Web corpus (csTenTen 12+17+19) [9], containing 14 billion tokens, morphologically annotated and lemmatized using Majka;
- for Croatian, the MaCoCu Croatian Web corpus v2 (2021–2022) [2], containing 2.7 billion tokens, morphologically annotated using RFTagger and lemmatized using CST Lemmatiser.

Both corpora are accessible in Sketch Engine [7].

Tables 1 and 2 provide an overview of the relevant available morphological categories for nouns in each language/corpus, along with the possible values and their representations in the tagsets.

In practice, available morphological information is somewhat limited by the actual tagger which was used in processing the corpus. For instance, Czech grammar distinguishes four genders for nouns (masculine animate, masculine inanimate, feminine, and neuter), which corresponds with the four values of the gender category in the tagset, which can all actually be found in the corpus

Table 1: Czech *ajka* part-of-speech tagset – noun categories

category	noun	k1
gender	masculine animate	gM
	masculine inanimate	gI
	feminine	gF
	neuter	gN
number	singular	nS
	plural	nP
case	nominative	c1
	genitive	c2
	...	

Table 2: Croatian *MULTEXT-East* part-of-speech tagset – noun categories

category	noun	N
gender	masculine	m
	feminine	f
	neuter	n
number	singular	s
	plural	p
case	nominative	n
	genitive	g
	...	
animate	no	n
	yes	y

data. On the other hand, while Croatian also distinguishes these four genders (represented in the tagset as two categories: gender and animate), the used tagger determined the animacy of masculine nouns only in the accusative case, meaning that data on animacy is not available for nouns in the genitive.

In addition to the explicitly tagged categories, we assumed there to be a dependency on the case of the initial letter of the lemma. In this way, we can distinguish multi-word proper names, such as personal names (e.g., *Božena Němcová*) and toponyms (e.g., *Frýdek Místek*), from noun clusters consisting of common nouns (e.g., *skupina lidí*, 'group of people'). In the Croatian tagset, there is a category called Type with the values *common* and *proper*, but we found it less reliable than looking at the initial letter of the lemma.

2 Methodology

After reviewing multiple hundreds of random two-noun clusters found in the corpora, we came to the conclusion that a sequence of **two nouns which are both in the genitive** is the single most productive source of ambiguity in the lemmatization of two-noun clusters. Therefore, we decided to focus our attention on disambiguating this type of noun clusters.

By observation, we established gender (including animacy) and the proper noun / common noun dichotomy as two possible triggers of the canonical form type *lemma + lemma* (as contrasted to the prevailing type *lemma + word*). We then searched the corpus for each possible combination of genders and initial-letter cases in each noun of a two-noun cluster. For each such concordance, we first inspected the first thirty lines in order to verify that the query was indeed correct and that there were no obvious errors in the tagging of the data. Since the data was not yet aggregated in any way at this time, this let us see also some less typical cases. Then we generated a frequency distribution of the found noun

clusters (which already brought the most frequent clusters to the top) and had these two-noun clusters displayed as each of the four combinations of word and lemma. Eventually, we inspected these lists and used our linguistic intuition to judge which combination of word and lemma is the appropriate canonical form type in each setting. We selected one actual noun cluster to represent each such decision in our analysis.

During the research, we targeted primarily correctly tagged Czech or Croatian words, respectively, and did not pay much attention to foreign words or obvious tagging errors. Yet, we still took into account the tagging of proper names, however arbitrary it might have seemed, e.g., *Marie Curie* tagged as feminine + masculine animate, or *Alfred Hitchcock* tagged as masculine animate + neuter. After all, there were not many other results for these combinations of gender and initial-letter case.

When we had accumulated the data from these corpus observations, we started grouping similar search parameters that resulted in same canonical form types.

3 Results

3.1 Czech Noun Clusters

The first group with observed particular behavior among Czech noun clusters were masculine animate nouns. If both nouns in a cluster were animate, regardless of letter case (e.g., *pana prezidenta*, 'of mister president'), or if at least one of them was animate and both started with a capital letter (e.g., *Kristiána Kodeta*, 'of Kristián Kodet'), both nouns need to be lemmatized (the *lemma + lemma* type) to produce the canonical form (*pan prezident*, *Kristián Kodet*).

Lemmatization of both nouns was also determined necessary for combinations of nouns that started with capital letters and had the same gender (e.g., *Boženy Němcové*, 'of Božena Němcová', becomes *Božena Němcová*; or *Frýdku Místku*, 'of Frýdek Místek', becomes *Frýdek Místek*). An exception to this rule is when both the nouns are neuter (e.g., *Muzea Těšínska*, 'of the Museum of the Těšín region'), in which case the second noun should preserve its original form after lemmatization (*lemma + word*, i.e. *Muzeum Těšínska*).

For all other combinations of gender and proper noun / common noun, the canonical form type *lemma + word* was found to be appropriate.

It should be noted, though, that the resulting lemmatization rules, illustrated in Table 3, do not take into account tagging errors in the corpus (e.g., *Rudé právo*, which is actually not a noun cluster), and also that there still exist some infrequent exceptions that are being evaluated incorrectly (e.g., *bratři Mrštíkové*, in which each noun is tagged with different animacy).

Judging by our findings, we believe that an important factor in the choice of correct canonical form is whether the noun cluster refers to a person. Such noun clusters are either multi-word proper names, or consist of a common noun that refers to a person, followed by (a part of) that person's name. The present

Table 3: Combinations of Czech noun categories yielding the *lemma + lemma* canonical form type (table cells provide illustrative examples of corresponding noun clusters)

	gM gM	gI gI	gF gF
common common	*pana prezidenta*	N/A	N/A
common proper	*malíře Muchy*	N/A	N/A
proper common	*Pepka Námořníka*[a]	N/A	N/A
proper proper	*Kristiána Kodeta*	*Frýdku Místku*	*Boženy Němcové*

rules allow only masculine noun clusters of this type to be identified, because their animacy is directly specified in their tags. Other "animate" noun clusters also appeared in our data, such as the feminine *kněžna Libuše*, but their animacy cannot be readily determined due to lack of corpus annotation.

3.2 Croatian Noun Clusters

Conducting similar research on the corpus of Croatian was complicated by the fact that the used tagger does not provide information on the animacy of nouns in the genitive.

The Croatian corpus was also five times smaller than the Czech corpus, resulting in a smaller variety of the contained proper names.

One type of proper names that stood out in the corpus were toponyms, often forming noun clusters when preceded by common nouns denoting their types (e.g., *grad Zagreb*, 'Zagreb city', or *otok Hvar*, 'Hvar island'). We found these cases frequent enough to be worth our attention. In order to be able to implement rules that match such accompanying common nouns, we compiled a list of the most frequent ones, which we further enlarged manually with nouns encountered within such clusters in Croatian-language maps (e.g., *rijeka*, 'river'; *vodotok*, 'watercourse'; *nizina*, 'lowland').

The prevailing canonical form type for Croatian two-noun clusters once again proved to be *lemma + word*. Observations made during corpus research led us to the decision to enforce lemmatization also on the second word of the two-noun cluster (i.e. to apply the *lemma + lemma* canonical form type) in the following three situations:

1. if both nouns are masculine, the first noun is a common noun, and the second noun is a proper noun (e.g., turning *pape Franje*, 'of pope Francis', into *papa Franjo*, 'pope Francis');
2. if the first noun is a common noun denoting a type of a toponym, and the second noun is a proper noun (e.g., turning *rijeke Cetine*, 'of Cetina river', into *rijeka Cetina*, 'Cetina river');
3. if both nouns are masculine proper nouns, and the first one does not denote a type of a toponym (e.g., turning *Djeda Mraza*, 'of Father Frost', into *Djed Mraz*, 'Father Frost').

Noun clusters that do not match any of the above situations are to be lemmatized using the canonical form type *lemma + word*, i.e. the grammatical case (genitive) of the second noun should be retained.

Please note that situations (1) and (3) above are intentionally limited to masculine nouns only, although similar rules likely apply also to nouns of other genders. However, we found the morphological tagging and the lemmatization of non-masculine proper nouns in the corpus to be frequently incorrect, forcing us to stay on the safe side by preserving the attested form of the second noun, thereby producing a valid canonical form candidate, although not always the appropriate one.

4 Conclusion and Future Work

Lemmatization rules for two-noun clusters, observed in corpora, have been implemented as corpus queries, expressed in the Corpus Query Language (CQL [6]), while making use of the *m4* macro processor to make orientation and maintainance easier, in accordance with common practice [3]. Due to their complexity and the limitations of the CQL, some of the designed rules had to be expressed using multiple corpus queries. Corpus queries along with their respective canonical form types, preceded by definitions of the used macros, make up a terminology extraction grammar that can be used for terminology extraction.

Our research on noun clusters has contributed to the development of new terminology extraction grammars for Czech and Croatian, already published and used for terminology extraction in Sketch Engine and OneClick Terms.

The work has not only improved the rendering of terms that are two-noun clusters, but the same improvements have been applied also to multiple types of longer terms, in which two-noun clusters appear as parts of more complex constructions (e.g., *zvláštního vyšetřovatele Chorozona*, 'of the special investigator Chorozon', where the two-noun cluster preceded by an adjective follows the same patterns as if standing alone). In the improved Czech terminology extraction grammar, which supports terms of up to six tokens, we have found use for the presented two-noun cluster rules within ten such complex term types.

While the present research has been limited to two-noun clusters, clusters of three nouns (e.g., Czech *databáze složek detergentů*, Croatian *faza podnošenja zahtjeva*), and perhaps even longer (e.g., Czech *areál rozšíření druhů dřevin*, Croatian *baza podataka otisaka prstiju*), should be considered in future research.

Known shortcomings of the presented lemmatization rules are mostly owed to limited or low-quality morphological annotations in corpus data. In some cases, we took the liberty of sacrificing very rare phenomena for simplicity of the resulting term grammar.

In this research, we focused on noun clusters in the genitive case, which are the biggest source of confusion during lemmatization. Some other grammatical cases, though, could create ambiguity too, namely the dative and the instrumental, as hinted by the typology of complements in [8].

Although early testing with end users has shown that the designed rules have indeed led to improvements in the lemmatization of noun clusters during terminology extraction, quantitative evaluation could provide a better idea on how well the present system performs and what space there is for further improvements.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Baisa, V., Michelfeit, J., Matuška, O.: Simplifying terminology extraction: OneClick Terms. The 9th International Corpus Linguistics Conference (2017), https://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2017/general/paper385.pdf
2. Bañón, M., Esplà-Gomis, M., Forcada, M.L., García-Romero, C., Kuzman, T., Ljubešić, N., Van Noord, R., Sempere, L.P., Ramírez-Sánchez, G., Rupnik, P., et al.: MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. In: 23rd Annual Conference of the European Association for Machine Translation, EAMT 2022. pp. 303–304. European Association for Machine Translation (2022), https://aclanthology.org/2022.eamt-1.41.pdf
3. Blahuš, M., Jakubíček, M., Cukr, M., Kovář, V., Suchomel, V.: Development of evidence-based grammars for terminology extraction in OneClick Terms. Electronic lexicography in the 21st century. Proceedings of the eLex 2023 conference pp. 650–662 (2023), https://www.sketchengine.eu/wp-content/uploads/Development-of-Evidence-Based-Grammars-for-Terminology-Extraction-in-OneClick-Terms.pdf
4. Caruso, Đ.Ž.: The syntax of nominal expressions in articleless languages: A split DP-analysis of Croatian nouns. Ph.D. thesis, Stuttgart, Universität Stuttgart, Diss., 2012 (2013), https://d-nb.info/103482290X/34
5. Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlỳ, P., Suchomel, V.: Finding terms in corpora for many languages with the Sketch Engine. In: Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics. pp. 53–56 (2014), https://www.sketchengine.eu/wp-content/uploads/Finding_Terms_2014.pdf
6. Jakubíček, M., Kilgarriff, A., McCarthy, D., Rychlý, P.: Fast syntactic searching in very large corpora for many languages. PACLIC pp. 741–47 (2010), https://www.sketchengine.eu/wp-content/uploads/Fast_syntactic_2010.pdf
7. Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V.: The Sketch Engine: ten years on. Lexicography **1**, 7–36 (2014), https://www.sketchengine.eu/wp-content/uploads/The_Sketch_Engine_2014.pdf
8. Rappaport, G.: The Slavic noun phrase. Position paper for Comparative Slavic Morphosyntax (1998), https://slaviccenters.duke.edu/sites/slaviccenters.duke.edu/files/media_items_files/10rappaport.original.pdf
9. Suchomel, V.: csTenTen17, a Recent Czech Web Corpus. In: Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2018. pp. 111–123 (2018), http://nlp.fi.muni.cz/raslan/raslan18.pdf