Legal Terms and Word Sketches
A Case Study

Eva Mráková and Karel Pala

Natural Language Processing Centre, Faculty of Informatics
Masaryk University, Brno, Czech Republic

Abstract. In this paper we describe an approach to the semiautomatic identification of legal terms in Czech texts. Our general goal is to offer supplementary tools for building dictionary of Czech law terms.

At first we used the VaDis partial parser for recognition of the complex nominal constructions in a legal text – the current version of the Penal Code of the Czech Republic. Headwords of the recognized structures are usually relevant legal terms. Then we employed the Sketch Engine to find Word Sketches of these relevant terms in a large corpus of the standard Czech Czes, because corpora of legal Czech texts are not available yet. In spite of the fact that we used common texts we obtained very good candidates for legal terms as a result.

We also discuss relations between VerbaLex frames of the selected group of Czech verbs with financial meaning that occur in legal texts and Word Sketches found for some of these verbs. It appears that the combination of the valency frames and Word Sketches provides good candidates for the legal terms as well.

The paper is conceived as a case study in which we describe collocational behaviour of the selected Czech noun phrases and also some verbs belonging to the financial domain.

1 Introduction

Previous work focused on legal term recognition in Czech texts is described in [1] and [2]. While the first paper concerns especially legal terms in the form of noun groups, the second has dealt with legal verbs and their valency frames. Here we present possible enhancements of the methods mentioned in these papers and we also suggest exploitation of other tools suitable for building a legal term dictionary. Another approach is described in [3] which mainly relies on manual processing of the legal texts when finding the legal terms.

2 Recognition of the Complex Nominal Constructions

The current version of the Penal Code of Czech Republic containing approx. 36,000 word forms had served as a data source for experiments described in [1]. These experiments were optimized for high speed and thus the partial

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2010, pp. 31–39, 2010. © Tribun EU 2010

Table 1. Examples of Complex Nominal Groups

complex nominal group	English equivalent
pachatel trestného činu	who committed a criminal act
spáchání trestného činu	committing a criminal act
pokus trestného činu	an attempt to commit a criminal act
znaky trestného činu	attributes of the criminal act
způsob provedení činu	way of the committing a criminal act
dokonání trestného činu	completing a criminal act
účastník trestného činu	participant of the criminal act
trestnost pokusu trestného činu	punishability of the attempt
doba spáchání činu	time of the committing a criminal act
povaha spáchaného činu	nature of the committed criminal act
stupeň nebezpečnosti činu	degree of the dangerousness of criminal act

parser VaDis [4] used for syntactic analysis was transformed into Perl regular expressions for this purpose. The result of the experiment were base forms of noun groups sorted according to their frequency and these groups should appear as entries in the legal electronic dictionary, which is in preparation [3].

We used the same data source but in comparison with the above mentioned approach we have been working with the original Prolog version of VaDis and focused on more complex nominal constructions and their hierarchical structuring. The analysis of the data took several minutes and it was acceptable without any need for an optimization. The recognized noun phrases were not sorted according to their frequency but were clustered according to their headwords. The headwords of big clusters were often essential legal terms like *čin (act)*, *trest (punishment)*, *pachatel (offender)*, *zákon (law)*, *sazba (penalty)*, *vazba (detention)*, *soud (court)*, *předpis (regulation)*, *opatření (measure)*, *následek (consequence)*, *škoda (damage)*, etc. Of course, there were also clusters whose headword's legal meaning requires wider context, such as *odpovědnost (responsibility*, *rozkaz (command)*, *společnost (society)*, *stát (government)* etc.

In each cluster we inspected the more complex nominal constructions. Table 1 shows some of such constructions for the cluster headword *čin (act)*.

It can be observed that (parts of) the complex nominal constructions are good candidates for legal "subterms" in the context of the particular headword, for instance, when describing the dictionary entry *(trestný) čin (criminal act)* we should describe (or refer to) the contextual words *pachatel (offender)*, *spáchání (commitment)*, *pokus (attempt)*, etc.

3 Word Sketches

Word sketches [5] are one page summaries of a word's grammatical and collocational behaviour and they represent a really helpful tool for building terminological dictionaries (and not only them). At the same time Word Sketch Engine produces information how firmly the individual members of the collocations are tied together – this is indicated by the salience parameter [6].

Based on the grammatical analysis, the Sketch Engine also produces a distributional *thesaurus* for a language, in which words occurring in similar settings, sharing the same collocates, are put together, and *sketch differences*, which specify similarities and differences between near-synonyms. The system is implemented in C++ and Python and designed for use over the web.

3.1 Obtaining Legal Terms through WSE

At the moment, unfortunately, we do not have any corpus of the Czech legal texts at our disposal which could be used as a direct source for the Sketch Engine. Instead, we have used big corpora of Czech common texts: SYN2000 [7] containing about 110 million tokens and Czes [8] containing 1,191,157,014 tokens (compiled at the NLP Centre FI MU and completed in 2009). It consists mostly of the newspaper texts downloaded from the Web. It is annotated grammatically with lemmas and POS-tags.

Although we did not use law texts as the source, the results of the Sketch Engine using the Czech sketch grammar [6] are interesting and quite promising. We explored Word Sketch tables for the headwords of clusters described in the previous section and we obtained several hundreds relevant legal terms from them.

Moreover, words in some Word Sketch tables form natural groups of legal terms. For instance, the Sketch table gen_1[1] for the headword *čin (act)* contains a reasonable classification of criminal acts. It is shown in Table 2.

Table 2. Sketch table gen_1 for *čin*

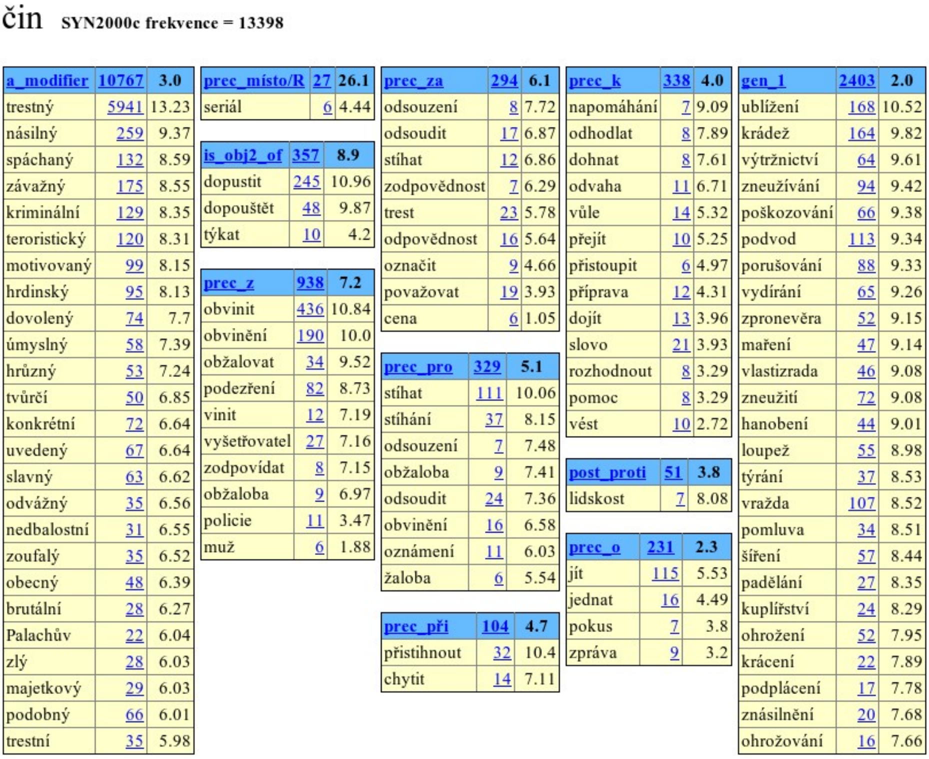
zpronevěra (defalcation)	padělání (forgery)
krádež (larceny)	hanobení (defamation)
poškozování (damaging)	podílnictví (shareholding)
loupež (robbery)	vražda (murder)
maření (obstruction)	zneužití (misaproppriation)
porušování (violation)	ohrožování (threatening)
zneužívání (misaproppriating)	zvýhodňování(privileging)
výtržnictví (disorderly behaviour)	ohrožení(emergency)
vydírání (extortion)	znásilnění (rape)
pomluva (slander)	podplácení (bribery)
zkrácení (reduction)	týrání (abuse)
zanedbání (negligence)	

Let us focuse on the headword *čin (act)* and its other Word Sketches. The word itself has a common meaning, not only the legal one. However, the Word Sketch Engine does not allow to process whole phrases like *trestný čin (criminal act)*, which is its "legal" specification. Despite of this, from general corpora consisting

[1] This is a very frequent collocation in Czech—it is a noun phrase consisting of the head noun and its dependent noun in genitive case, e. g. *pachatel trestného činu (the one who committed a criminal act – criminal, offender)*

of the common (e.g. newspaper) texts we obtained Word Sketch Tables with mainly legal terminology.

We have found almost 13,400 occurences of the *čin (act)* in SYN2000 and 88,500 ones in Czes. Corresponding Word Sketch tables contained more than one hundred words and more than two hundreds words respectively. A part of the Word Sketch Table of *čin (act)* obtained from SYN2000 is shown in Figure 1.



čin SYN2000c frekvence = 13398

a_modifier	10767	3.0
trestný	5941	13.23
násilný	259	9.37
spáchaný	132	8.59
závažný	175	8.55
kriminální	129	8.35
teroristický	120	8.31
motivovaný	99	8.15
hrdinský	95	8.13
dovolený	74	7.7
úmyslný	58	7.39
hrůzný	53	7.24
tvůrčí	50	6.85
konkrétní	72	6.64
uvedený	67	6.64
slavný	63	6.62
odvážný	35	6.56
nedbalostní	31	6.55
zoufalý	35	6.52
obecný	48	6.39
brutální	28	6.27
Palachův	22	6.04
zlý	28	6.03
majetkový	29	6.03
podobný	66	6.01
trestní	35	5.98

prec_místo/R	27	26.1
seriál	6	4.44

is_obj2_of	357	8.9
dopustit	245	10.96
dopouštět	48	9.87
týkat	10	4.2

prec_z	938	7.2
obvinit	436	10.84
obvinění	190	10.0
obžalovat	34	9.52
podezření	82	8.73
vinit	12	7.19
vyšetřovatel	27	7.16
zodpovídat	8	7.15
obžaloba	9	6.97
policie	11	3.47
muž	6	1.88

prec_za	294	6.1
odsouzení	8	7.72
odsoudit	17	6.87
stíhat	12	6.86
zodpovědnost	7	6.29
trest	23	5.78
odpovědnost	16	5.64
označit	9	4.66
považovat	19	3.93
cena	6	1.05

prec_pro	329	5.1
stíhat	111	10.06
stíhání	37	8.15
odsouzení	7	7.48
obžaloba	9	7.41
odsoudit	24	7.36
obvinění	16	6.58
oznámení	11	6.03
žaloba	6	5.54

prec_při	104	4.7
přistihnout	32	10.4
chytit	14	7.11

prec_k	338	4.0
napomáhání	7	9.09
odhodlat	8	7.89
dohnat	8	7.61
odvaha	11	6.71
vůle	14	5.32
přejít	10	5.25
přistoupit	6	4.97
příprava	12	4.31
dojít	13	3.96
slovo	21	3.93
rozhodnout	8	3.29
pomoc	8	3.29
vést	10	2.72

post_proti	51	3.8
lidskost	7	8.08

prec_o	231	2.3
jít	115	5.53
jednat	16	4.49
pokus	7	3.8
zpráva	9	3.2

gen_1	2403	2.0
ublížení	168	10.52
krádež	164	9.82
výtržnictví	64	9.61
zneužívání	94	9.42
poškozování	66	9.38
podvod	113	9.34
porušování	88	9.33
vydírání	65	9.26
zpronevěra	52	9.15
maření	47	9.14
vlastizrada	46	9.08
zneužití	72	9.08
hanobení	44	9.01
loupež	55	8.98
týrání	37	8.53
vražda	107	8.52
pomluva	34	8.51
šíření	57	8.44
padělání	27	8.35
kuplířství	24	8.29
ohrožení	52	7.95
krácení	22	7.89
podplácení	17	7.78
znásilnění	20	7.68
ohrožování	16	7.66

Fig. 1. A part of the WS Table of *čin (act)* in SYN2000

3.2 Terminological Verbs

Together with already mentioned gen_1 table we also obtained Sketch tables containing several verbs with legal meaning. Some of them are listed in Table 3. In [1] a group of verbs collected from legal texts was investigated. They represented a mixture of the various common verbs and also some legal ones. In comparison with them the verbs obtained now from the Sketch tables are actually terminological verbs with legal meaning. Some of them could be straightforwardly used as entries in a dictionary of legal terms.

Legal verbs were explored in [2] and they were added to the lexical database VerbaLex [9]. In this way the Verbalex was extended with a reasonable number of the legal verbs. However, it is still possible to find candidates of legal verbs for further extension of VerbaLex in our Word Sketch tables (e.g. *promlčet(be time-barred)*, *překvalifikovat (change qualification)*, *prošetřovat (investigate)*, *zpochybňovat (question)*,. . .).

Table 3. Legal verbs from word sketch tables

spáchat (commit)	dopouštět se (perpetrate)
páchat (commit)	odsoudit (condemn, sentence)
vyšetřovat (investigate)	potrestat (punish)
překvalifikovat (change qualification)	přiznat (confess)
prošetřovat (investigate)	postihovat (affect)
vykonat (perform)	prokázat (prove)
zmařit (thwart)	ohlásit (announce)
ospravedlňovat (justify)	uprchnout (escape)
stíhat (prosecute)	zodpovídat (be responsible)
objasňovat (explain)	zadržet (arrest, detain)
promlčet (be time-barred)	napravit (amend)
zpochybňovat (question)	litovat (regret)

In Sketch tables we, of course, find also other parts of speech but they usually do not contain any data relevant for legal terminology (prepositions, particles, etc.). However, there is one more interesting Sketch table that should be mentioned—the one with adjectives. While adjectives are not typical dictionary entries, some of them should be explained at least in a hierarchical context of the headword *čin (act)* (e.g. *úmyslný (deliberate)*, *nedbalostní (caused by negligence)*, *násilný (violent)*, *protiprávní (illegal))*.

3.3 Verbs with Financial Meaning

Verbs explored in [2] also include a group of verbs occuring in legal text and belonging to the financial domain. While the verbs mentioned above were processed by the WSE here we decided to have a look at the verbs in whose complex valency frames the argument labeled as EXT(sum:1) occurs. Then we explored their frequencies in the corpus Czes (see Table 4 on the next page).

First, it has to be remarked that the verbs in the list fall into small subgroups containing semantically close items – they are either aspect pairs or even triples, if iteratives are considered. We will not deal with the pairs perfective : imperfective here, the category of aspect belongs to the area of morphology in Czech.

It can be observed that the differences in the frequencies of the particular verbs in the table are significant. It is not difficult to conclude that the less frequent verbs in the list display specialized terminological meanings, for

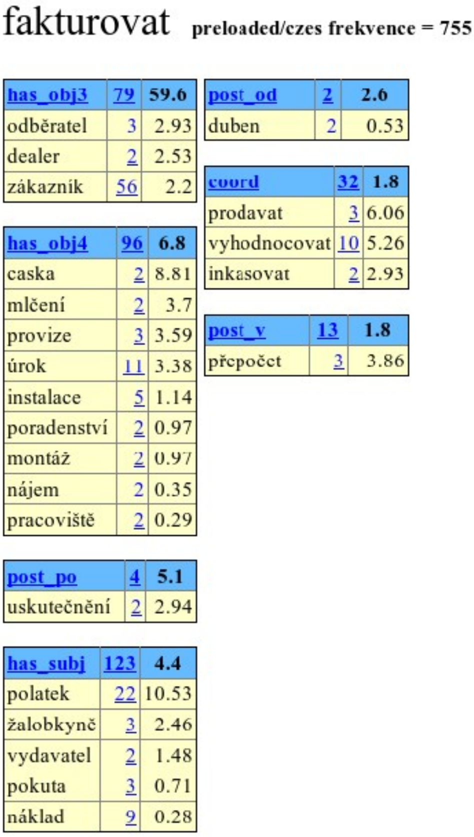
Table 4. Examples of Financial Verbs

Verb	frequency in Czes
alokovat (allocate)	670
realokovat (reallocate)	13
danit (tax)	45,374
zdanit (tax)	3,291
dodanit (pay up the tax)	117
dodaňovat (pay up the tax)	20
dlužit (owe, have a debt)	11,773
vydlužit (take on loan)	13
fakturovat (invoice)	755
vyfakturovat (invoice)	135
financovat (finance)	18,598
dofinancovat (finance up)	219
předfinancovat (prefinance)	19
počítat (calculate, compute)	116,673
spočítat (calculate, compute)	13,663
tarifikovat (tariff)	23
tarifovat (tariff)	12
validovat (validate)	65
valorizovat (valorise)	591
platit (pay)	290,253
zaplatit (pay up)	148,683
splatit (pay off)	10,787
proplatit (pay out, cash)	2,070
proclít (clear through customs)	119
vyclít (clear through customs)	8
vydražit (auction off)	2,465
vydražovat (be auctioning)	11

instance *vyclít (clear through customs)* with the frequency 8 or *předfinancovat (prefinance)* with 19. We are aware that the frequency cannot serve as the only convincing indicator of the terminological status of these verbs – more detailed evaluation would be needed. In any case, for the verbs in the Table 4 we can say that the ones with the frequency lower than 1,000 can be reliably considered terminological.

4 Verbalex Valency Frames and Legal Terms

In this section we will briefly touch the relation between complex valency frames of the financial verbs as they can be found in VerbaLex and their corresponding Word Sketches obtained from the corpus Czes. The assumption is that the semantic labels of the verb's arguments such as AG(person:1|institution:1) or EXT(sum:1) match reasonably with the concrete nouns that appear in the Word Sketch tables of the respective verbs.



fakturovat preloaded/czes frekvence = 755

has_obj3	79	59.6
odběratel	3	2.93
dealer	2	2.53
zákazník	56	2.2

has_obj4	96	6.8
caska	2	8.81
mlčení	2	3.7
provize	3	3.59
úrok	11	3.38
instalace	5	1.14
poradenství	2	0.97
montáž	2	0.97
nájem	2	0.35
pracoviště	2	0.29

post_po	4	5.1
uskutečnění	2	2.94

has_subj	123	4.4
polatek	22	10.53
žalobkynč	3	2.46
vydavatel	2	1.48
pokuta	3	0.71
náklad	9	0.28

post_od	2	2.6
duben	2	0.53

coord	32	1.8
prodavat	3	6.06
vyhodnocovat	10	5.26
inkasovat	2	2.93

post_v	13	1.8
přcpočct	3	3.86

Fig. 2. WS of *fakturovat (invoice)* in Czes

Take, for instance, the complex valency frame for the verb synset *vy/fakturovat (invoice)* capturing its financial meaning:

1: fakturovat$_{n1}$, zaúčtovat$_{n1}$, zaúčtovávat$_{n1}$, naúčtovat$_{n1}$, naúčtovávat$_{n1}$
AG<person:1>$^{obl}_{kdo1}$ VERB^{obl} REC<person:1|institut.:1>$^{opt}_{komu3}$ ART<goods:1>|ACT<act:2>$^{opt}_{za+co4}$ EXT<sum:1>$^{obl}_{co4}$
–example: advokátka si fakturovala za své služby desetitisíce korun (impf)
–example: obsluha hostovi zaúčtovala stopadesátikorunový poplatek (pf)
–example: číšník zákazníkovi naúčtuje špatnou cenu (pf)
–synonym:
–use: prim
–reflexivity: obj_dat

In the valency frame of the verb *fakturovat (invoice)* we find the arguments labeled as AG<person:1>, REC<person:1|institut.:1>, ART<goods:1>, ACT<act:2>, EXT<sum:1>. The question thus is whether they have real counterparts (tokens) in the Word Sketch (Table 2). The simple manual comparison shows that the answer is positive and that nouns found in the respective corpus sentences semantically agree with what is predicted by the argument labels in the valency frame. We think that it is not not necessary to go into details here but the next step should consist in an attempt to formulate a formal procedure that would perform exactly this.

It has to be remarked that in corpus sentences we observe that some of the arguments are frequently expressed by personal pronouns (mostly classified as subjects and objects), thus we should be able recognize that e.g. personal pronouns *já, ty, on, mu (I, you, he, him)* match with labels like AG(person:). These cases have to be handled by a procedure defined just for this purpose. The next phenomenon that we must deal with are passive verbs forms, which, as we can see in the corpus, because of their transitivity transform the order of the arguments and their surface valencies, i.e. they (in Czech) substitute the accusative case with nominative or instrumental with nominative. The verbs in VerbaLex contain information about their transitivity or intransitivity but we need to formulate transformation rules which will do this automatically when it is needed – the passive verb forms then will serve as a trigger.

Another phenomenon that plays a relevant role in this context are anaphora relations and their resolution. The frequency of the personal pronouns that function as antecedents in anaphoras is quite high, for instance the frequency of *já (I)* in the corpus Czes is 1,981,248, the frequency of *on (he)* is 5,726,584, thus role of the anaphorical relations cannot be neglected. Unfortunately, the present versions of the algorithms handling the resolution of the anaphorical relations in Czech are not successful enough for the indicated task. It also has to be taken into account that processing of the personal pronouns by the Sketch Engine is still at its beginning. The handling of the demonstrative pronouns is also relevant in this respect and we are afraid that it is even more difficult task.

5 Conclusions

In this paper we have discussed some possible techniques for semiautomatic finding the terminological entries that could be used in building the Czech legal term dictionary. We have investigated the behaviour of the noun candidates of legal terms using the Word Sketch Engine and can conclude that the obtained results are promising though, at the moment, we cannot offer a complete quantitative evaluation. We also explored some verbs with financial meaning – for them we found that their frequencies in the corpus Czes convincingly prove their terminological nature. At the end we have briefly dealt with the relation between complex valency frames of the financial verbs as they can be found in VerbaLex and their corresponding Word Sketches obtained from the corpus Czes. This comparison shows that in this way it is possible to obtain more detailed information about the meaning of the verbs belonging to the financial domain but not only for them. These observations can be generalized also for the verbs from other domains.

Acknowledgements

This work has been partly supported by the Ministry of Education of CR within the Center of basic research LC536 and in the National Research Programme II project 2C06009 and by the Czech Science Foundation under the projects P401/10/0792 and 407/07/0679.

References

1. Pala, K., Rychlý, P., Šmerk, P.: Automatic Identification of Legal Terms in Czech Law Texts. In: Semantic Processing of Legal Texts, Berlin, Springer (2010) 83–94.
2. Pala, K., Mráková, E.: Verb Valency Frames in Czech Legal Texts. In: Proceedings of the RASLAN 2009 Workshop, Brno, Masaryk University (2009).
3. Cvrček, F.: Právní informatika (Legal Informatics). Publisher A. Čeněk, Plzeň (2010).
4. Mráková, E.: Partial Syntactic Analysis (of Czech). Ph.D. thesis, Faculty of Informatics, Masaryk University, Brno (2002) (in Czech).
5. Kilgarriff, A., Rychlý, P., Smrž, P., Tugwell, D.: The Sketch Engine. In: Proceedings of the Eleventh EURALEX International Congress, Lorient, France: Universite de Bretagne-Sud (2004) 105–116.
6. Pala, K., Rychlý, P.: A Case Study in Word Sketches – Czech Verb *vidět (see)*. In: A Way with Words: Recent Advances in Lexical Theory and Analysis (A Festschrift for Patrick Hanks), Menha Publishers (2010) 187–198.
7. Ústav Českého národního korpusu: Korpus SYN2000. `http://ucnk.ff.cuni.cz/syn2000.php` (2000).
8. Natural Language Processing Centre, FI MU: Czes Corpus. `http://corpora.fi.muni.cz/ske/auth/` (2009).
9. Horák, A., Hlaváčková, D.: VerbaLex – New Comprehensive Lexicon of Verb Valencies for Czech. In: Computer Treatment of Slavic and East European Languages, Third International Seminar, Bratislava, VEDA (2005) 107–115.