The Design of Czech Lexical Database

František Čermák
Institute of Czech National Corpus, Philosophical Faculty, Prague, Czech Republic
Frantisek.Cermak@ff.cuni.cz

Jana Klímová
Institute of Czech Language, Czech Academy of Science, Prague, Czech Republic
Jana.Klimova@ff.cuni.cz

Karel Pala
Dept. of Information Technologies,
Faculty of Informatics, Masaryk University Brno, Czech Republic
pala@fi.muni.cz

Vladimír Petkevič
Institute of Theoretical and Computational Linguistics
Philosophical Faculty, Charles University, Prague, Czech Republic
Vladimír.Petkevič@ff.cuni.cz

1. Introduction

The aim of the paper is to present a conception of Czech Lexical Database (CLD) that should later become a basis for the new representative Czech dictionary. The main purpose of this enterprise is to build a representative Czech Lexical Database that would serve as a source of lexical information and also as a partial knowledge representation in various NLP applications (Ingria, Boguraev, Pustejovsky, 1992, p.341-365)[1]. The basic units in CLD can be either single lemmata like *dům (house)* or standard collocations as e.g. *vysoká škola* (*university*). The assumed size of the designed CLD is approximately 50 000 entries and we set as our primary task to concentrate on Czech verbs as much as possible, i.e. the number of the described verbs should be about 20 000 (40 % of CLD size, when the estimated number verbs in Czech is about 40 000 items). This is based on the fact that the verbs represent the main relational element in natural languages around which the other elements, mostly nouns, are centered.

2. The basic structure of CLD

It can be described by a DTD (most probably designed in XML) that will consist of the following parts (fields, see e.g. Faber, Usón 1999, p.20)[2]:

a1) <phonological (phonetic) information> about the sound structure of the expressions constituting a given entry. This in fact means that we will be trying to develop the (parallel) speech database for Czech that would form a data collection for building the algorithms able to process speech signals, as e.g. speech recognition, synthesis and encoding, as well as recognition and verification of the speaker. The speech database data will be appropriately included in the lexical database. Some interesting tasks have to be solved in this respect: particularly, additional word forms will have to be generated by the speech synthesis module since it would be virtually impossible to account for all forms of all words in the lexical database: there are approximately 5–5.5 million word forms in Czech).

a2) <morphological information> about the structure of the entry (for Czech) – it will yield the information about POS and all the respective grammatical categories associated with that POS plus the information about the basic segmentation. For nouns this can be captured by <a code of the respective inflectional paradigm> since we assume that a morphological analyzer/generator AJKA will be integrated into CLD (Sedláček, 1999)[3] that would offer the morphological information on demand

[1] Ingria R, Boguraev B, Pustejovsky J 1992 Dictionary/Lexicon. in Encyclopedia of Artifical Intelligence (ed. by Shapiro S. C.), New York, John Wiley, pp.341-365.

[2] Faber P, Usón R, M 1999 Constructing a Lexicon of English Verbs. Berlin – New York, de Gruyter.

[3] Sedláček R 1999 Morfologický analyzátor pro češtinu (Morphological analyser for Czech). Diploma Thesis, Faculty of Informatics, Brno.

(dynamically). For verbs this typically includes 8 categories (attributes): <negation>, <person>, <number>, <tense>, <mode>, <voice>, <aspect> and <gender>. Their values would be accessed dynamically through the <inflectional paradigm code of the verb>. To get this information the morphological analyzer/generator will be used in a similar way as for nouns. Also word formation information will be included here in a subfield and it should show the relevant formally justified links between the respective entries including their semantic consequences, cross POS relations like the direction of derivation (as in *práce → pracovat (work → to work)*). This poses a task how to formulate the word derivation rules as formally as possible (see below Klímová, Pala, 2000, p.987-991),

a3) <senses1...n> where for each sense the following should be given:

a3.2) <semantic features> that can be associated with an entry – possibly based on EuroWordNet Top Ontology and hypero/hyponymy hierarchies (trees) or their parts (subtrees or clusters) (Vossen, 1999)[4]. It is to be examined how large parts of the trees or subtrees can be employed – we estimate that the plausible number of the nodes used here may be about 5,

a3.3 <descriptions using genus proximum (hypero/hyponymic relationships)> and distinguishers (differentia specifica), will be given typically for the entries containing nouns. In fact the genus proximum definitions can be viewed as subsets of the hypero/hyponymy trees where only two nodes are considered. The distinguishers represent a sort of problem: in our view they quite successfully resist to the attempts to formalize them. This can be demonstrated by the fact that the particular dictionaries differ most in the way in which they treat the distinguishers – there is no general agreement as to what distinguishers should or should not be selected and included in the particular entries.

a3.4 <semantic classes> – for verbs the genus proximum definitions may not work as reliably as for nouns or they may be used reliably only for a small number of them, therefore we suggest here to indicate a semantic class a verb belongs to. In this respect we are preparing a semantic classification of Czech verbs similar to Levin‘s (Levin, 1995)[5] though in Czech this task seems to appear more complicated because of the category of aspect (thanks to this Czech verbs regularly occur in pairs). On the other hand, it is also obvious that the semantic classes of verbs are closely related to the valency frames (verb frames) and we set as our task to reflect these links in the database as well.

a3.1 <synset> that can be found for a given lexical unit (entry, lemma), possibly in WordNet fashion. The reason for having synsets follows from the fact that the relation of synonymy (and antonymy) can serve as one of few relatively reliable ways of the characterization of meaning

a4) <syntactic information> about the combinatorial properties of the entry and the expressions that are related to it. It is obvious that typically the syntactic properties of the given item are strongly related to the particular sense of the entry and they distinguish it from the other senses. The information given in this field will be captured through <valency frames> for all the POS where it makes sense, i.e. for verbs, nouns, adjectives, numerals and also some adverbs. It is evident that in this respect we have to distinguish syntactic (or superficial) valency frames that in Czech will include the combinatorial information about the morphological cases (there are seven of them in Czech) and semantic (or deep) valency frames containing the necessary information about the semantic cases (roles) that are expressed by the morphological cases. The notation linking syntactic and semantic valencies is indicated in the examples below, however we take it as preliminary since the final inventory of deep cases for Czech has not been established yet (see e.g. Fillmore and Atkins, 1998, p. 417-423)[6] .

That is not all, in our view it is also very useful to include the particular lexical information in the valency frames. Typically, it is not enough to know just the respective values of morphological (superficial) cases but their lexical "cast" as well, see e.g. the vital difference between two accusatives occurring in *držet v ruce knihu (hold a book in the hand)* and *držet tvar (to mantain, keep the shape)*. It can be objected that the semantic valencies should capture these differences in the senses but for practical NLP applications it certainly appears to be practical to have this kind of information in CLD in an explicit form.

a5) <local contexts>, i.e. contexts typical of the given entry, e.g. *hezká dívka (pretty girl),* or *šikovný chlapec (smart boy)* etc., they will be obtained from the corpus.

a6) <examples or typical uses>, e.g. *držet knihu v ruce (to hold a book in the hand), otočit hlavu (to turn the head),* we should get them from the corpus texts,

a7) <collocations> with appropriate subcategorizations, i.e. an attempt has to be made to find a

[4] Vossen P et al. 1999 Final Report on EuroWordNet-2, 2D041. CD ROM, v.1, Amsterdam, University of Amsterdam.

[5] Levin B 1995 English Verb Classes and Alternations. Chicago, The University of Chicago Press.

[6] Fillmore Ch, Atkins B 1998 FrameNet and Lexicographic Relevance, in Proceedings of the First National Conference on Language Resources and Evaluation. (eds. Rubio A, Gallardo N, Castro R, Tejada A) , vol. 1, Paris, ELRA, pp. 417-423.

semantic classification of the collocations. We will not go into details here but it can be shown that it would be very useful to have e.g. the verbal collocations classified in accordance with the semantic classes of verbs mentioned above. Similar technique can be applied to the noun collocations as well but we are aware of the fact that this task will require a lot of corpus data and their laborious analysis.

a8) <pragmatic information> – more structured information about the register and stylistic properties of the item including the regional information and other data, however, we would like to cover only the basic types of this information,

a9) <origin> – i.e. short etymological information related to the item,

a10) <logtype> – this field will include the information about the logical type of the item based on the Transparent Intensional Logic (TIL) (Materna 2000[7], Pala 2000, p.109-114)[8]. In TIL the types are built on the ramified theory of types which, we hope, may lead to formally more consistent semantic representations of NL expressions. This together with hierarchic hypero/hyponymic structures will enable us to use CLD also as a part of the knowledge representation systems. We would like to try to establish the relations between EuroWordNet Top Ontology and Type Ontology as defined within TIL. This should yield more precise and less arbitrary semantic classifications for the semantic hierarchies, semantic relations and semantic features as well though on the other hand we are aware that this enterprise may also lead to some problems, e.g. it may be feasible only for some entries or parts of speech (verbs, nouns, adjectives, adverbs).

a12) <encyclopedic information> – can be included in CLD where it would be useful or even necessary for possible NLP applications, this may hold e.g. for the entries that are related to the information technologies. The question is whether we should look for a kind of knowledge representation language that would enable to represent the encyclopedic information or to approach the problem pragmatically and to follow the present encyclopedic resources in their current form. In the examples below we give only some arbitrary explanations but in the beginning it appears to be more reasonable to follow the latter path.

3. Resources for CLD

Thanks to favourable situation with regard to Czech National Corpus (CNK, at Charles University in Prague, Faculty of Arts – (abbrev. FF UK) and corpus ESO (at Masaryk University in Brno, Faculty of Informatics – (abbrev. FI MU) we assume that the building of CLD can be based mainly on the CNK and ESO data. Other resources would be used as well, particularly the two existing Czech dictionaries:

1) large Dictionary of Literary Czech (SSJČ, 1960)[9],
2) and smaller Dictionary of Written Czech (SSČ, 1984)[10].

Additional resources will have to be sought as well, particularly other existing dictionaries, especially the terminological ones. We are also aware of the fact that a well founded reader program has to be established sooner or later that should be later closely connected with the preparation of the New Czech Dictionary.

4. Tools

Thanks to the interesting results in the NLP research being performed both at Charles University (Institute of Czech National Corpus, Institute of Formal and Applied Linguistics, Institute of Theoretical and Computational Linguistics) in Prague and Masaryk University in Brno (Faculty of Informatics, NLP Laboratory) we have at our disposal a basic set of tools that can be used for building CLD.

Particularly, we will take advantage of the morphological analyzer AJKA (mentioned above), parsers (DIS and GT, Žáčková, Popelínský, Nepil, 2000, p.219-225, Horák, Smrž, 2000, p.43-50)[11], desambiguators (Oliva, Petkevič et al., 2000, p. 3-8)[12], corpus manager and graphical interface using

[7] Materna P 2000 Type-theoretical analysis as a preparation of analyzing expressions of a natural language. Prague - Brno, Faculty of Informatics MU, manuscript, pp.110.

[8] Pala K 2000 Word Senses and Semantic Representations - Can We Have Both? in Proceedings of TSD 2000, Berlin, Springer Verlag, pp.109-114.

[9] Slovník spisovného jazyka českého (Dictionary of Written Czech Language) 1960, Praha, Academia.

[10] Slovník spisovné češtiny (Dictionary of Literary Czech) 1984, Praha, Academia.

[11] Žáčková E, Popelínský L, Nepil M 2000 Recognition and Tagging of Compound Verb Groups in Czech. in Proceedings of CoNLL-2000 and LLL-2000, Lisbon, ACL New Brunswick, pp.219-225.Horák A, Smrž P 2000 Large Scale Parsing of Czech, in Proceedings of Efficiency in Large-Scale Parsing Systems Workshop, COLING'2000, Universitat des Saarlandes, Saarbruecken, pp.43-50.

[12] Oliva K, Petkevič V et al. 2000 The Linguistic Basis of a Rule-Based Tagger of Czech. in

client-server architecture (Rychlý, 2000)[13], dictionary editor and a browser based on XML format that can handle any dictionary converted into XML. The modified version of the browser can be also used for processing any WordNet data in a way that is now possible with Polaris database (Vossen, 1999).

Other tools include various conversion programs, programs for corpus maintenance and corpus preparation, heuristic programs for obtaining valency frames from corpus texts, Czech morphological database and programs for automatic word derivation that would capture the word derivation chains like *učit (to teach) – učení (teaching) – učitel (teacher) – učitelka (she-teacher) – učený (scholarly) – učenec (scholar) – výuka (tuition, lesson)* etc. It has to be decided whether these data should be included in CLD directly or rather whether they would be obtained dynamically from the morphological module (Klímová, Pala, 2000, p.987-991)[14]. We touched this question above when discussing the morphological information about the entries. In the examples given in the next section we have not tried to show the word formation semantic information because we hope that it can be obtained dynamically as the output of the word formation engine that is presently being built for Czech at NLP Lab. at Faculty of Informatics MU.

5. Conclusions

In this short contribution we have presented the underlying assumptions from which building Czech Lexical Database can start. We are aware of the fact that a number of the discussed points will have to be elaborated more deeply and systematically to obtain fully applicable results. This explains the fact that the examples of the entries are in several points rather tentative skeletons than full and complete entries. However, it is our hope that the described techniques, resources and tools will allow us to reach our goal.

5.1 An example[15]

An example of the entry for the Czech verb *držet (hold)* follows:

```
<entry <držet>>
<mf <k5> <eA> <nS> <pI> <t*> <m*> <aI> <paradigm code>>
<sense:1 <def: uchopovat rukou, mít v ruce>>
        <val: <ag<k1člověkc1>> <obj<k3něcoc4> <ins<<k7vc6> <k1rucec6>>>
        <synset: <uchopovat> <mít v ruce> >
        <sfeat: <činnost> <...>  %(WN-like)
        <semclass: <3.1> %(verbs of holding and keeping)
         <contexts: <držet dveře, d. pistoli v ruce>   >  %(examples from corpus)
            <collocation: <> <>  >  %(+ semantic class of a collocation)
              <style: written>
                <etym:  …              >
                  <logtype: vztah mezi dvěma individui, relation-in-intension between two individuals>
                         <encyc: ruka je část lidského těla nebo robota> >

<sense:2 <def: být pevnou součástí jiného objektu>>
        <val <obj<k1c4> <ins<<k7c6> <k1c6>>>
        <synset: <být fixován> <být upevněn> >
        <sfeat <...> <mero...> %(WN-like)
         <semclass: <3.2>   %(verbs of holding )
          <contexts: <omítka drží> >  %(examples from corpus)
            <collocations: <hřebík drží> <>  > %(+ semantic class of a collocation)
              <style: written>
                <etym: …              >
                  <logtype: vlastnost individua, property of an individual >>
```

Proceedings of TSD 2000, Berlin, Springer Verlag, pp.3-8.

[13] Rychlý P 2000 Korpusové manažery a jejich efektivní implementace (Corpus Managers and their Effective Implementation). Ph.D. Dissertation, Brno, Faculty of Informatics MU.

[14] Klímová J, Pala K 2000 Application of WordNet ILR in Czech Word-formation. In Proceedings of LREC Conference, Athens, ELRA, pp.987-991.

[15] Note that in the examples English is used to describe the fields, however, we doubt whether it makes sense to translate Czech lexical data within the fields. We decided not give the translations and to offer the necessary explanations during the poster session when discussing the particular points.

<encyc: upevnění objektu se provádí lepením, zatlučením> >

<sense:3 <def: zachovávat tvar>>
<val <obj<k1*kloubouk*c1>> <forma<k1*tvar*c4> >>
<synset: <zůstávat v pevném tvaru> <neměnit formu> >
<sfeat <...> <...> > %(WN-like)
<semclass: <3.3> %(verbs of maintaining and keeping shape)
<contexts: <vlasy drží fazónu> > %(examples from corpus)
<collocations: <puky drží> <> > (roztřídit podle typů)
<style: written>
<etym: …   >
<logtype: vlastnost individua, property of an individual >>
<encyc: platí pro objekty jako šaty, vlasy> >

<sense:4 <def: zachovávat polohu>>
<val <ag<k1*člověk*c1>> <obj<k1*tělo*c4>> <mod<k6xM*rovně*> >>
<synset: <být ve stejné poloze> <neměnit polohu> >
<sfeat <...> <...> > %(WN-like)
<semclass: <4.1> %(verbs of keeping a position)
<contexts: <držet tělo vzpřímeně> > %(examples from corpus)
<collocations: <držet hlavu nad vodou> <> > %(+ semantic class of a collocation)
<style: written>
<etym: …   >
<logtype: vlastnost individua, property of an individual >>
<encyc: platí pro polohu lidského těla> >

<sense:6 <def: vlastnit půdu>>
<val <ag<k1*vlastník*c1>> <obj<k1majetekc4> > >
<synset: <vlastnit půdu, spravovat majetek> >
<sfeat <...> <...> > %(WN-like)
<semclass: <2.1> %(verbs of possesion)
<contexts: <držet zahradu> > %(examples from corpus)
<collocations: <držet byt> <držet dům> > %(+ semantic class of a collocation)
<style: written>
<etym: ….   >
<logtype: <vztah mezi individui, relation-in-intension>>
<encyc: vlastníkem je člověk, objektem nemovitost> >

<sense:7 <def: >>
<val <ag<k1*člověk*c1>> <obj<k1*zvířata*c4> >>
<synset: <pěstovat> <chovat> >
<sfeat <...> <...> > %(WN-like)
<semclass: <5.1> %(verbs of growing)
<contexts: <babička drží slepice> > %(examples from corpus)
<collocations: <držet dobytek> <> > %(+ semantic class of a collocation)
<style: written>
<etym: …   >
<logtype: <vztah mezi individui, relation-in-intension>>
<encyc: objekty jsou zvířata> >

<sense:8 <def: rezervovat>>
<val <ag<k1*člověk*c1>> <obj<k1*člověk*c3>> <loc<k1*místo*c4>> >
<synset: <rezervovat, obsadit> >
<sfeat <...> <...> > %(WN-like)
<semclass: <6.2> %(verbs of reservation and booking)
<contexts: <drží nám místo v pořadí> > %(examples from corpus)
<collocations: <držet komu místo> <> > %(+ semantic class of a collocation)
<style: written>
<etym: …   >
<logtype: <vztah mezi individui, relation-in-intension> >

<encyc: platí pro místo v dopr.prostředku nebo v seznamu> >

<sense:8 <def: oblíbit si koho, preferovat koho>>
<val <ag<k1*člověk*c1>> <obj <<k7*na*> <k1*člověka*c4>> | <věc <<k1*tenis*c4> > >
<synset: < oblíbit si koho, preferovat koho > >
<sfeat <...> <...> > %(WN-like)
<semclass: <7.2.> %(verbs of emotional attitudes, liking)
<contexts: <on na ni drží> > %(examples from corpus)
<collocations: <držet na sestru> <> > %(+ semantic class of a collocation)
<style: written>
<etym: … >
<logtype: <vztah mezi individui, relation-in-intension> >
<encyc: platí pro místo lidi nebo člověka a nějaký oblíbený objekt (peníze)> >

<sense:8 <def: držet s kým>>
<val <ag<k1*člověk*c1>> <obj<k7*s*c7> k1*člověk*c7>> >
<synset: <držet s kým, držet spolu, být s kým v partě> >
<sfeat <...> <...> > %(WN-like)
<semclass: <8.4> %(verbs of social grouping)
<contexts: <držet partu> > %(examples from corpus)
<collocations: <držet s komunisty> <> > %(+ semantic class of a collocation)
<style: written>
<etym: … >
<logtype: <vztah mezi individui, relation-in-intension> >
<encyc: platí pro místo v dopr.prostředku nebo v seznamu> >

<sense:9 <def: uchopovat rukou, mít v ruce>>
<val: <ag<k1*člověk*c1>> <obj<k3*něco*c4> <ins<<k7*v*c6> <k1*ruce*c6>>>
<synset: <uchopovat> <mít v ruce> >
<sfeat: <činnost> <...> %(WN-like)
<semclass: <3.1> %(verbs of holding and keeping)
<contexts: <držet dveře, d. pistoli v ruce> > %(examples from corpus)
<collocation: <> <> > %(+ semantic class of a collocation)
<style: written>
<etym: … >
<logtype: vztah mezi dvěma individui, relation-in-intension between two individuals>
<encyc: ruka je část lidského těla nebo robota> >

5. 2 Example of the entry for Czech noun *hlava* (*head*):

<entry <hlava>>
<mf <paradigm code: 47a> <<k1> <gF> <nS> <c1>> >
<sense:1 <defgenprox: část těla>>
<difspec: >
<val <<k1*hlava*c1> <k1*člověka*c2>> <<k2*psí*c1> <k1*hlava*c1>> >
<synset: <kebule> <palice> <šiška>>
<sfeat <holo: tělo> <mero: tělo, oči, nos, tváře, ústa> > %WN-like
<contexts: <lidská hlava> > %(examples from corpus)
<collocations: <c1> <c2> …<cn> > %(+ semantic class of a collocation)
<styl: written>
<etym: ... >
<logtype: property of an individual>
<encyc: platí pro hlavu člověka, zvířete nebo robota> >

<sense:2 <def: rozum, mysl>>
<difspec: >
<val <<k1*hlava*c1> <k7*na*c4> <k1*fyziku*c4>> <<k2*chytrá*c1> <k1*hlava*c1>> >

```
        <synset: <mysl> <vědomí> >
        <sfeat <hyper: abstr>   > %WN-like
        <contexts: <chytrá hlava> >  %(examples from corpus)
         <collocation: <c1> <c2> …<cn>  >   %(+ semantic class of a collocation)
               <styl: written>
                    <etym: …   >
                        <logtype: vlastnost individua, property of an individual>
                             <encyc: sídlo myšlení, inteligence u lidí> >

<sense:3 <defgenprox: šéf, vedoucí skupiny>>
               <difspec:  >
        <val <<k1hlavac1> <k1podnikuc2> > >
        <synset: <boss> <náčelník> >
        <sfeat <hyper: hum> <hypo: podřízený>   > %WN-like
        <contexts: <hlava mafie> >  %(examples from corpus)
          <collocations: <hlava rodiny> <c2> …<cn> > %(+ semantic class of a collocation)
             <style: <written>
               <etym: …              >
                  <logtype: vlastnost individua, property of an individual>
                           <encyc: v hierarchické organizaci, firmě, vládě> >

<sense:4 <defgenprox:  přední část předmětu>>
               <difspec:  >
        <val <<k1hlavac1>  <k1kolac2>> >
        <synset: <výstupek> <> >
        <sfeat: <hyper: objekt> <hypo: >  > %WN-like
        <contexts: <hlava šroubu> >  %(examples from corpus)
         <collocations: <c1> <c2> …<cn> >   %(+ semantic class of a collocation)
             <style: written>
               <etym: …     >
                  <logtype: vlastnost individua, property of an individual >
                     <encyc: technical component> >

<sense:5 <defgenprox:  jednotka textu >>
               <difspec:  >
        <val <<k1hlavac1> <k1zákonac2 >>>
        <synset: <oddíl, kapitola v knize> <> >
        <sfeat: <holo: kniha> <mero: paragraf>  > %WN-like
        <contexts: <hlava 22> > %(examples from corpus)
           <collocations: <> <>  >  %(+ semantic class of a collocation)
            <style: written>
               <etym: …              >
                 <logtype: vlastnost individua, property of an individual >
                    <encyc: obsah knihy se člení na jednotky – kapitoly, hlavy >
```