Notes on Compiling a Corpus-Based Dictionary*

František Čermák, *Institute of the Czech National Corpus, Faculty of Philosophy, Charles University, Prague, Czech Republic (frantisek.cermak@ff.cuni.cz)*

Abstract: On the basis of sample analysis of a Czech adjective, a definition based on the data drawn from the Czech National Corpus (cf. Čermák and Schmiedtová 2003) is gradually compiled and finally offered, pointing at the drawbacks of definitions found in traditional dictionaries. Steps undertaken here are then generalized and used, in an ordered sequence (similar to a work-flow ordering), as topics, briefly discussed in the second part to which lexicographers of monolingual dictionaries should pay attention. These are supplemented by additional remarks and caveats useful in the compilation of a dictionary. Thus, a brief survey of some of the major steps of dictionary compilation is presented here, supplemented by the original Czech data, analyzed in their raw, though semiotically classified form.

Keywords: MONOLINGUAL DICTIONARIES, CORPUS LEXICOGRAPHY, SYNTAGMATICS AND PARADIGMATICS IN DICTIONARIES, DICTIONARY ENTRY, TYPES OF LEMMA, PRAGMATICS, TREATMENT OF MEANING, POLYSEMY, CZECH

Opsomming: Aantekeninge oor die samestelling van 'n korpusgebaseerde woordeboek. Op grond van 'n steekproefontleding van 'n Tsjeggiese adjektief, word 'n definisie gebaseer op data ontleen aan die Tsjeggiese Nasionale Korpus (cf. Čermák en Schmiedtová 2003) geleidelik saamgestel en uiteindelik aangebied wat wys op die gebreke van definisies aangetref in tradisionele woordeboeke. Stappe wat hier onderneem word, word dan veralgemeen en gebruik in 'n geordende reeks (soortgelyk aan 'n werkvloeiordening), as onderwerpe, kortliks bespreek in die tweede deel, waaraan leksikograwe van eentalige woordeboeke aandag behoort te gee. Hulle word aangevul deur bykomende opmerkings en waarskuwings wat nuttig is vir die samestelling van 'n woordeboek. Op dié manier word 'n kort oorsig van sommige van die hoofstappe van woordeboeksamestelling hier aangebied, aangevul deur die oorspronklike Tsjeggiese data, ontleed in hul onbewerkte, alhoewel semioties geklassifiseerde vorm.

Sleutelwoorde: EENTALIGE WOORDEBOEKE, KORPUSLEKSIKOGRAFIE, SINTAGMATIEK EN PARADIGMATIEK IN WOORDEBOEKE, WOORDEBOEKINSKRYWING, SOORTE LEMMAS, PRAGMATIEK, BEHANDELING VAN BETEKENIS, POLISEMIE, TSJEGGIES

* This article is an edited version of a plenary address delivered at the conference on 'Dictionaries, More than Words', which took place at the Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia, 6 February 2009.

Lexikos 20 (AFRILEX-reeks/series 20: 2010): 559-579

Introductory Remarks

The following notes discuss some of the problems and issues encountered during the compilation of a monolingual dictionary. After some preliminary remarks, these notes are split into two main parts. In Part I (2–4), an example and discussion of an analysis of corpus data (drawn from Czech) is offered, resulting in a lexical profile of a word. In Part II (5–8), building on the previous part, a commented list of some of the main aspects and principles of the dictionary-making process is presented. References to most of the points raised are to be found in the literature at the end.

Before any work can begin, a series of decisions laying down the ground rules must be made. Some of these are quite straightforward, others may be more difficult. The major ones may briefly be listed as follows:

Firstly, (a) resources have to be decided on. These include a large balanced (representative) corpus, in addition to some secondary resources, if available. Here it is necessary to be rather wary of the Internet, which is not always as generally useful as is often presumed.

Next, (b) the type of dictionary has to be decided. In the present context, such a dictionary should have, among others, the attributes monolingual, large, synchronic, representative and descriptive (not prescriptive).

Connected with this, though not necessarily dependent on it, is to make a decision about (c) the target users, for example the general public as well as a specialized public.

Finally, taking account of the shortcomings of existing dictionaries and the advantages of having a corpus, a major decision should be made, to be applied in various forms, about (d) the main concerns and orientation of the proposed new dictionary. Three general aspects should be stressed: syntagmatics, usage and context. This implies that attention will be paid to all relevant variants of words and phrases in a language, a fact that is often underestimated or even disliked by prescriptive normalisers and codifiers of a language.

Some attention should be given to (e) software. Here, one of the obvious candidates available today is represented by TshwaneLex. The choice should be made with a view to possible re-use of the data, expansion, redesign, additional products, etc.

PART I

1. Data and Treatment

Any dictionary aims — or should aim — at a true mapping of the meaning behind words, though modern dictionaries include mapping of their use, too, which is made possible by modern corpus data. A large dictionary should respect these data as best it can: it will basically be a corpus-driven product. A major new problem for lexicographers is the profusion of data that was not

available until only a few years ago. For this reason, ways and means are still being sought how to best handle such abundant data. Special and constant problems are identification of the meaning behind corpus forms, splitting the corpus data into manageable groups, interrelationships and elegant and efficient description. By way of analysis of a given lemma, it will, hopefully, be possible to point to some central problems of description and explanation of the meaning. Differences between the traditional approach and the corpus approach can best be illustrated if both the description of a given lemma, taken from an existing dictionary, and a description compiled on the basis of corpus analysis are compared in some detail. As a rule in theoretical models of data analysis, attention is paid almost exclusively to verbs and their salient formal and semantic features, following the preoccupations of syntacticians, while almost no attention is paid to nouns, the largest of the word classes. The latter are seriously in need of more detailed inspection. In addition, some attention must also be paid to adjectives, which stand between the two poles.

2. Traditional Approach and a Critique

First a case may be considered where a pair of related traditional dictionaries (SSČ and SSJČ) and corpus evidence differ widely. The example selected is the Czech polysemous adjective *měkký* (which corresponds roughly to English *soft*). The portrait of this adjective in dictionaries is rather brief and looks simple and straightforward:

(1) *poddávající se, málo odolávající tlaku* (yielding or giving way to pressure)

(2) *vzbuzující (na pohled n. na poslech) dojem jemnosti* (evoking an impression of fineness or tenderness (to the eye or ear))

(3) *podléhající snadno citu, citlivý, soucitný, povolný* (succumbing to feelings, sensitive, compassionate, compliant)

An inspection of the corpus data (3 549 occurrences in SYN2000 of some 100 million words) — or a sample of it that is presumed to be sufficient — yields a rather different picture. The very first impression one gets is that something is seriously wrong with the dictionary definitions. They do not exhaust the data; they use a problematic metalanguage (employing, among other things, synonyms); and they dissect the semantic continuum in a way that is odd, if not wrong. Specifically, they omit a number of analytical criteria that suggest themselves on inspection of the corpus data. Some of these may briefly be outlined. The first few examples, backed by the corpus, deal with the three meanings given in the dictionaries.

(a) *Pressure* (*tlak* in definition 1). Questions one must ask here include the following:
Is it physical or psychological (abstract) pressure that is meant?

Does *měkká voda* (soft water) yield or give way to pressure? (Hardly!)
What about *měkká norma* (soft norm)? (It does not fit.)
Does definition 1 imply that a soft object may be crumpled, cut, sawed or torn apart? (Not under normal circumstances.)

(b) Definition 2 relies heavily on the near synonym *jemnost* (fineness, tenderness). This is problematic. The adjective *jemný* is given five meanings in the dictionaries (**1** having a smooth surface; **2** graceful/delicate; **3** having a small degree of a quality perceived by the senses; **4** distinguishing exact details; **5** having a specifically high quality). These five meanings are not sufficient to cover the meanings of collocations found in the corpus such as *měkké pohyby* (soft movements), *měkký hlas* (soft voice), *měkká stupnice* (minor key, in music), and *měkké i* (soft i, in orthography). In none of these collocations is the synonym *jemný* correct. Thus, the reference to *jemný* is misleading or useless.

(c) *Podléhající citu* etc. (feelings, definition 3). Here too, it is difficult to fit this definition to existing collocations, for example *měkký člověk* (soft-hearted man), *měkká povaha* (conciliatory nature), etc. They are different and difficult to describe in this way.

(d) Next, there is a multitude of examples, illustrated by corpus collocations, that do not fit the definitions either, e.g. *měkký horský vzduch* (soft air), *měkká ekonomika* ('soft economy'), *měkká koncepce* ('soft conception'), *měkké dřevo* (soft wood), *měkké lyže* ('soft skis'), *měkká pornografie* (soft pornography), *měkká radiace* (soft radiation), etc.

3. Some Principles of Corpus Data Analysis

To get to the bottom of the maze of facts that lie behind this adjective and not to leave out anything relevant, a comprehensive and representative (if not exhaustive) concordance must be compiled and analysed. The analysis must be based on random samples, whose number and size will depend on the type and complexity of the lemma being analysed. The analysis that seems to be relevant in most cases consists of a number of steps, mostly manual and rarely simple, always starting from features found in the data (steps 3–5). However, it is necessary to start (steps 3.1 and 3.2) by singling out and setting aside cases that would otherwise complicate the analysis.

3.1 Idioms and Phrasemes

Without going into detail, all of these may be identified on the basis of a paradigmatic or syntagmatic anomaly, which is either semantic or formal in nature. Here, only a few cases are eligible. These include the expressions *mít měkké srdce* (be soft-hearted) and *být měkký na někoho* (to be soft — i.e. not strict — on sb).

Additionally, it should be noted that, although no examples are found in the case of *měkký*, fixed expressions and stereotypical phrases, including catch phrases and proverbs, fall under this heading too.

3.2 Multi-Word Terms

Leaving aside instances of specific terminological meanings of single-word lemmas, which are typical of nouns, there is, in the case of this adjective, some terminology that consists of multi-word terms, such as *měkká voda* (soft water), *měkká droga* (soft drug), *měkká radiace* (soft radiation), *měkká pornographie* (soft pornography), and *měkký konec řádky* (soft end-of-line return).

After this, the gist of the analysis is concentrated in three steps (3.3–3.5).

3.3 Determination of Function

Of the three main adjectival functions, namely (a) attributive only, (b) predicative only and (c) both (majority of adjectives), it turns out that all uses of *měkký* fit into the last type only. Hence no specific functional description is necessary here, though other adjectives may have more specific functionality. Obviously, each word class has one or more specific functions, distinct from the other.

3.4 Semiotic Classification

This largely depends on the part of speech. It is basically pragmatic and corresponds to particular needs. In the present case, it seems sufficient to classify all the nouns qualified by the adjective *měkký* into five broad classes according to the type of denotation of the noun that they modify, namely:

(a) man (humanus, H), *obchodník* (*byl*) *měkký,*

(b) animal (animalis, An), *krávy jsou měkké,*

(c) (concrete) thing (res concreta, K), *řízek* (*byl*) *měkký*,

(d) (abstract) thing or abstract (res abstracta, A), *měkká atmosféra*, and

(e) (place (locus, L), –.

In some cases, it may be useful to identify a sixth class, namely:

(f) metaphorical use (M), *měkká politika* (literally, soft politics).

This is discussed further under point 3.8 below. Most uses of the adjective *měkký* in the corpus data under inspection fall into (c) and (d).

Only when this analysis is complete, is it viable to look, within these broad classes, for any further markers and features, which may be very important but do not seem to be so general.

3.5 Formal markers

These include any relevant information that the form signals. A desideratum here, though difficult to attain, is to do automatically as much identification as possible of at least the following formal features:

(a) valency, most prominent with verbs though not limited to them,

(b) special position or formal use, and

(c) specific frequent collocates.

While a single valency to be found here is restricted to the idiom mentioned above, no postpositive uses of the adjective *měkký* were found (though some adjectives are so used). Neither was any specific uses of *měkký* with negatives or other special constructions encountered. However, there are some frequent and obvious cases of *měkký* found collocating with *být* (to be), which should be duly noted.

At least two more systematic criteria should be applied in any analysis of corpus data for a lexical item. These are the paradigmatic set membership of the item and its frequency.

3.6 Set Membership in a Collocational Paradigm

By this, the whole range of regular collocations of the item is meant, with the exception of idioms and multi-word terms, though these are closely related. It is no paradox to view a set of collocations, i.e. syntagmatic feature, as a collocational paradigm (one or more). This has not yet been done systematically in any dictionary. However, it gives vital information about the possible uses of the item in text, so it is crucial to mention this kind of information. For practical purposes, this becomes of greatest importance in those cases where the collocational set (paradigm) is comparatively small, restricted to only a few members (i.e. a closed paradigm set). Although no such restricted collocational sets are to be found in the case of *měkký,* the point can easily be illustrated by a different word, the Czech adverb *dokořán*, which is translated as 'fully' in bilingual dictionaries. The fact is, however, this word collocates with only six other words (*otevřít, být, nechat, zůstat; okno, dveře*, i.e. open, be, leave, remain; door, window). It is, then, far more important to give these six collocations in the dictionary, not trying to determine the meaning of *dokořán* at any cost, for this is not easy to specify (in some cases it corresponds to English 'ajar'). In an attempt to find the meaning in this case, generalizing over a mere six occurrences is linguistically problematic: there may not be a sufficient analogy here. A sufficient analogy is a prerequisite for any judgements about the meaning of a lexical item and its type.

3.7 Frequency

It is almost impossible to overestimate the importance of this feature, which is now so well documented in the corpus, but which users, until now, have had no access to. It helps in many ways, not least by indicating which meaning should be recorded as the first in the dictionary.

Before continuing, two more remarks of a general nature must be made.

3.8 Paradigmatic–Syntagmatic

Though a good corpus may offer many different types of information, handling this information may be somewhat idiosyncratic, depending on the type of dictionary. It is evident that new, corpus-based dictionaries should aim to redress the age-old imbalance in information offered previously. As a generalization, it may be said that these dictionaries, because of the limited supply of data and their main purpose, have largely been skewed towards the paradigmatic aspect, emphasizing classifications of various sorts and determining memberships in classes set up by lexicographers.

With modern corpora, however, it is possible, for the first time ever, to offer syntagmatic information in dictionaries as well, indicating vital information about the usage of words in real texts. In lexicography, this amounts to two things primarily, valency and collocations. Though formal valency (such as the case forms required by prepositions) may not be difficult to pin down and should and can be determined for all word classes (not only verbs, such as *depend on, abstain from*), collocations still present a problem. It is not so much a matter of their exact theoretical determination — though linguists take widely different positions on this — but rather a matter of practical selection from the vast quantities of corpus data.

One of the problems created by the profusion of data in modern corpora is that one is pushed, by means of various statistical association measures (such as log-likelihood or MI score) towards what is *typical* only, being offered little or no information about marginal, infrequent and, perhaps, untypical uses, which a large dictionary should record or illustrate too. To view marginal collocations as a limitless string of exceptional, figurative or metaphorical uses is hardly a solution. Instead, potentiality of use should be considered here and instances of isolated marginal use should be double-checked against other resources. No doubt, in some cases, such collocations will turn out to be no isolated or figurative uses, but newly emerging types of standard meaning.

3.9 Pragmatic Uses

Finally, pragmatic uses should be identified and a specific semiotic approach devised. What effect does a particular expression have on the reader or listener,

and under what circumstances? A major feature here is evaluative use, which, as it happens, is often of a negative nature.

4. Lexical Profile of the Adjective *měkký* (soft)

The analysis based on the points raised and briefly explained above, has produced a different profile of the lexeme *měkký* from the one with which was started (Čermák 2007). This profile is shown in what follows (though it could, based on different emphases, take other shapes, too). Even the best dictionaries differ widely from any corpus-based profile. So far, few dictionaries have been based on corpus data, and none in Czech. Obviously, a profile such as the one below, if applied in a printed dictionary, would have to be collapsed into the dictionary's description format. It would, however, be expected to preserve all the distinctions found in the corpus and mentioned here (above and below) and to be made clear for the user. The latter point imposes the constraint of a limited metalanguage vocabulary. It is evident that the syntagmatic aspect is made prominent here, especially in the subsenses (a), (b), (c), etc. A sample of *měkký* (soft) that has been analysed is given in the Appendix. The lexical profile, originally compiled in Czech (see Appendix 1), is given here in English for the benefit of a wider readership.

1. **ABILITY and EFFECT (of a concrete object) that is physical for the agent (animate):** *under the influence of pressure or force, easy to shape, cut, saw or fold; elastic and quite resilient*
   (a) matter, material, product: *having a smooth surface, pleasant to touch*
   (b) physical object, product: *rounded, not angular*
   (c) fruit: *very ripe*
   (d) meal: *prepared, cooked and ready for eating*

2. **EFFECT (of a concrete or abstract object) that is physical, especially acoustic, visual or tactile, for the receiver (inanimate or animate):** *having a pleasant quality including a fine effect or contrast rather than being sharp or pronounced*
   (a) voice, sound: *quiet and delicate*
   (b) rain etc.: *not strong, neither severe*
   (c) contact, fall, blow: *not violent or intensive*
   (d) consonant: *pronounced as fricative*

3. **EFFECT (of an abstract object) that is psychological for the receiver (animate):** *being sympathetic, benevolent or even compassionate, and sometimes slightly exaggerated*
   (a) words, language: *not stern neither angry, conciliatory*
   (b) a human being in their conduct or expression: *conciliatory in politics or irresolute*
   (c) norm, judicial decision: *not severe, not principled or consistent*

4. **EFFECT of a concrete or abstract object that is different (from that in 1–3) on the receiver (animate):**
   (a) alcoholic drink or other intoxicating substance: *having a weak effect*
   (b) market, currency, goods: *losing value*
   (c) water: *without minerals (and unsuitable, among others, for shaving)*
   (d) drug: *not addictive*
   (e) radiation: *weakly penetrating*
   (f) pornography: *suggestive, rather than explicitly erotic*

Further criteria could be introduced to make the overall picture more detailed, such as distinguishing cases where the concrete and abstract are collapsed. This all depends on the degree of granularity that the lexicographer wants to achieve. Naturally, the more detailed the description gets, the less transparent and organized for the user it becomes. The fourth major class, which is complementary to the first three, covers residual types of meaning and usage, and is often terminological and metaphorical; it could easily be expanded into separate categories.

PART II

Notes on Some Stages and Types of the Lexicographer's Work

5. General and Theoretical Issues

Drawing to some extent on the preceding part, which was more practical in nature, some generalizations will be mentioned in this part. The following notes, more theoretical and often very short, do not aspire to be a systematic and full survey of the problems that lexicographers deal with (see, for example, Hartmann and James (1998), Atkins and Rundell (2008) and Hanks (2009, Forthcoming)).

5.1 The Basic Resource: The Corpus

A good and balanced corpus is today essential for the compilation of a dictionary, but it is sometimes necessary to consult additional resources (such as those mentioned in 5.2), either because more information is required or because corroboration of corpus data is needed.

5.1.1 Word List and Frequencies

— A frequency list of words and lemmas is very useful for many purposes, e.g. for determining the likely complexity of an entry.

— Frequency information should be given for all lemmas.

— The list should include all variants found in the corpus, ordered by frequency.

— All members of a closed class should be included (e.g. names of colours).

5.1.2 Selection for Analysis

— When selecting a sample from the corpus for analysis, it is important to avoid one text genre only, wherever possible, and at all events to avoid relying on a single source, which would be too skewed and likely to result in distortion.

5.1.3 Concordances

— The choice of random samples is necessary, if the data for a particular lexical item is too big.

— A manageable selection in a concordance has its limitations, though ordering it may help to overcome some of these, for example to find formal valency markers, collocations, etc.

— Filters can help in making a further selection, if these are available in the corpus browser.

— Statistical measures may offer additional help in decision making, especially with respect to collocations.

5.1.4 Additional Corpus Tools

— Other tools are available, such as *Word Sketches*, though they do not help in decisions about peripheral phenomena.

5.2 Additional Resources

Should the corpus data not be sufficient, then targeted excerpts or even inquiry through distributed questionnaires in special cases might be necessary. (It very much depends on the corpus composition.)

The Internet is not to be trusted as a source of data, in many cases being skewed and full of hiatuses. Its worst performance is probably in the domain of authentic spoken language and dialogue.

5.3 Types of Lemma (Dictionary Macrostructure)

At least four types of lemma/entry should be distinguished, namely:

— Single-word lemmas: most entries; no grouping is preferable.

— Multi-word lemmas: idioms and terms, problems of selection and identification.

— Technical apparatus: abbreviations, cross-references, etc.

— Specialized types of entry may also be envisaged, for example prefixes and suffixes or suppletion forms having a different alphabetical order (English *went, go*).

5.4 The Entry: Some of its Features (Dictionary Microstructure)

In what follows, the single-word lemma will be specifically commented on.

5.4.1 Form

Form includes a number of familiar items whose treatment depends on the dictionary policy. Here, only a brief summary will be given:

— Lemma, variants (a true description of forms that have actually been recorded, not prescriptive, otherwise it could lead to a never-ending selection).

— Grammar (endings, reference to tables, etc.).

— Pronunciation (differential only, some foreign words).

5.4.2 Style, Register

The dictionary should reflect real usage (in contrast to stylistic theories, which are usually far from the world of real language). Labelling should be kept to a minimum and terms must be designated. As register tends to change rather rapidly, any labelling should be reviewed at the end of the compilation.

5.4.3 Additional Features (optional)

With a large dictionary many options open up, which cannot be given much thought and scope in lesser ones. These may include special sections on:

— Frequency (in some simplified form).

— Synonymy (though this should never be a substitute for meaning definition).

— Etymology.

— Special usage notes (mostly pragmatic, perhaps also historical, including notes on differences between variants too).

5.5 Meaning

Rendering a satisfactory description of an item's meaning is the most important goal of any general dictionary. Only a few basic principles will be mentioned:

— Meaning and use are inseparable because meaning can only be deduced from attested use.

— Meaning can be deduced only from real and sufficient contexts of use.

— Each definition should be self-sufficient, not relying on outside information.

— Each definition should be worded sufficiently, so that it does not fit other entries, i.e. it should be unique.

— Definitions should be based on real data and should hold for all significant occurrences of the form.

5.5.1 Types of Meaning

A distinction should be made between the meaning of (a) terms (see 3.2, 6.2) and (b) standard lexemes; both being further distinguished from (c) pragmatics (such as evaluative function).

5.5.2 Definitions

Except for the COBUILD type of definition, most approaches are basically variations of the mainstream. Some of the salient principles are as follows:

— The basic, classical approach is based on the genus proximum + differentia specifica dichotomy, i.e. where possible. In today's terms, this boils down to a closest hypernym and a set of specific necessary features.

— Ostensive, deictical definition is useful (if available), using showing and pointing (though indirectly in most cases) to outside objects and phenomena the word is related to. This may include pictures, charts, etc.

— Relational definitions hold for derivatives, but the semantic relations are not always mechanical and additive. This is a frequent source of misinformation as the derivatives hardly ever reflect the base exactly, e.g. between a noun and a related adjective.

— Often, it is useful to give typical referential nouns (for adjectives) or the type of subject, object, etc. (for verbs). This is directly linked with collocations and other syntagmatic information.

— Function (of grammar words, etc.) is not meaning, nor can it be related to other specific lexemes (by way of collocations, etc.).

— Since function is theory-dependent (e.g. conjunctions and particles depend on a theory of syntax and pragmatics), the relevant theory must be stated explicitly in advance, at least by reference to a particular framework.

5.5.3 Polysemy

Polysemy, universal in language in all of its frequent lexemes, traditionally causes difficulties for the lexicographer, there being no consensus as to how to deal with it (but see the suggestions above, put forward on the basis of the analysis of *měkký*). At least the following general points can be made:

— The meaning and its parts/senses should be related to form wherever possible (i.e. syntactic use, valency, collocability).

— Discrimination between common usage and terminological phraseology is necessary.

5.5.4 Other Semantic Features

These may be viewed as largely (though not invariably) complementary, including:

— Synonyms may superficially seem to be useful, though ideally the users might expect comment about differences between a synonym and the lemma.

— Opposites (not just plain antonyms) are essential if available, helping the users and orienting them in the lexicon system.

— Hypernyms (not necessarily only the immediate ones) are essential and in fact no definition is possible without them.

5.6 Principles of Meaning Definition

A number of specific principles can be mentioned that relate to the description of meaning. Though commonplace, perhaps, these are worth giving here for they should always be kept in mind. Consider at least the following:

— The unknown (and rare) should be explained in terms of the known (and common). There is an advantage in having a specified metalanguage (e.g. the Longman restricted defining vocabulary of 3 000 common and frequent words), though this has not yet been tried for a large dictionary.

— Context and usage is the main arbiter for the meaning of a word often standing in sharp contrast to preconceived ideas.

— There is no standard size of context to be given in examples; it depends on the nature of the lemma.

— Each definition should be equivalent in its form and wording to the relevant part of speech, enabling a possible substitution in text (for autosemantic/lexical words). Here, a broad substitution test (substituting the definition for the lemma in relevant contexts) may often be helpful.

Nevertheless, the use of paraphrase in the definition should be unambiguous. An alternative is the COBUILD full-sentence type of definition.

— The definition must not be circular (no defining by mutual synonyms is a solution or description).

— Opposites and contrasting words, if there are any, must be mentioned as these are important links to a complementary lexeme.

— All examples should correspond to the definitions given and should not substitute those parts of it that are not mentioned.

— There is no such thing as a specific isolated meaning: The solution is either to find more examples to make it a standard meaning or to declare the combination to be an idiom. The old idea of exception, preserved in grammar perhaps, can be dissolved into either solution indicated above.

— As much as possible must be fitted into the definition, avoiding metaphorical meanings, perhaps by a double-layered approach (i.e. giving a main meaning plus secondary meanings to each sense).

— The possibility must be considered of giving a (simplified) scientific definition of terms versus standard definition (e.g. defining *salt* as 'NaCl, sodium chloride', as well as 'a white crystalline substance used for seasoning or preserving food').

— Collocational restrictions must be observed: If a lemma is found to collocate with a severely restricted class of collocates only, this must either be explicitly stated or the class must be viewed as a set of 'fixed' collocations (idioms) and the lemma must be taken out of the list as not being in use independently.

— (Morphological) forms, occurring in specific collocations usually, often have a specific meaning, not applicable to the whole lemma, hence they may require special treatment in a section of the dictionary article or in an independent lemma.

— Extended, mostly metaphorical cases of use should be carefully selected, if intended for inclusion, especially with regard to showing possibilities of (current or future) expansion of standard meanings that have been recorded, as an indication of the potentiality of the language.

6. Idioms and Terminology (single and multi-word lexemes).

As both idioms and terminology are a matter of a much more complicated and different type of lexicography (see, for example, Čermák 2007), only a couple of general principles may be mentioned here.

6.1 Idioms

These should be given sufficient definitions, including information about use, the classes of users, and the circumstances under which they are used.

— All idioms should remain unrelated to numbered meanings of a single-word lemma and should receive special treatment, including specification of their pragmatic function.

— At least some idioms/phrasemes could be independent entries.

— Many idioms are pragmatic, specifically evaluative and this information should be explicitly given.

— The problem of their alphabetisation should have a simple and systematic solution (e.g., for word classes: first noun, then adjective, then verb, etc.).

6.2 Terms

Constituting the largest part of any natural language (including numerous multi-word expressions), these should always be defined in consultation with experts, who should also assist in the selection of technical terminology.

— In many cases, terms should be given both an encyclopedic and lexicographic definition; the latter may be shorter.

— There are no self-evident criteria for the choice of terms. Some combination of expert advice and corpus frequencies is needed.

— It may be desirable to distinguish between the terminological and common use of lemmas. (See the discussion of *salt* above.)

Finally, it may be useful to mention briefly some practical issues regarding the whole process of dictionary compilation.

7. Technical Aspects of Compilation

There are very many technical aspects of dictionary compilation. Only two of them will be mentioned here.

— A preliminary database could be useful. If available and preannotated, it will save time during the compilation process, although regrettably there is a danger that it might overlook new data if the corpus is growing during the period of compilation.

— Useful tools include ready-made templates (for data split into homogeneous classes) and a style guide (mainly specifying the sequence of steps to be taken and editorial policy decisions).

7.1 Preparation

Once a word-list is available it is advisable to:

— Split entries into homogenous categories, such as parts of speech and their subclasses: This safeguards homogeneity of compilation. But not all words are easily classified in this way.

— Compile an average and a medium-size entry as a pilot exercise: This will provide valuable experience and a basis for modulation of principles. Obtaining a first idea of what an average entry will be like usually serves as a basis for planning (though the conditions specified in any plan are rarely met and fulfilled in all details).

7.2 Further Steps

Before starting in earnest on compiling the dictionary, it is useful to ensure that the data is as homogeneous as possible. This, among others, means:

— Selecting and extracting idioms and other multi-word units for special treatment (see 3.1 and 3.2).

— Identifying pragmatic words and expressions (i.e. those related to society, addressing the basic question 'How does their use affect people?'): Special descriptions accounting for social use (and abuse) and effect are necessary here.

— Creating lexical profiles as a starting point: A useful tool is *Word Sketches*, but only for the core usage of each word.

7.3 Technical Aids

— Statistical association measures such as MI-score and t-score indicate salient combinations and their types. Sometimes even simple bi-/trigrams are helpful, too. On the other hand, no single association measure yields all the collocations that might be of interest.

— Collocations are scalar, ranging from typical to rare and untypical. (A policy is therefore needed to decide how far to go with the inclusion of these.)

7.4 Control Mechanisms

Some control mechanisms are necessary, designed in particular to ensure that

— the same types of entry are handled similarly throughout (for all members of a class), and

— formal mechanisms (such as punctuation and spacing), references, etc. are styled consistently throughout.

8. Open Questions

A number of open questions remain, depending on the specific procedures used. The following at least may be mentioned:

— Maintaining links with an open corpus, specifically when in need of further or new data.

— Drawing the line, i.e. finding a cut-off point between collocations that are quoted and those omitted.

— Including dialect forms and information about them.

Acknowledgement

The author's thanks go to Patrick Hanks for his suggestions in helping to improve the final text.

References

Dictionaries

SSČ = *Slovník spisovné češtiny pro školu a veřejnost*, ed. J. Filipec, F. Daneš, J. Machač, V. Mejstřík, 2. vyd. 1994 (1. vyd. 1978). Praha: Academia.

SSJČ = *Slovník spisovného jazyka českého* 1960–1971. Praha: NČSAV/Academia.

Other Literature

Atkins, B.T. Sue and M. Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Čermák, František. 2007. Polysémie a kolokace: případ adjektiva měkký. Čermák, F. and M. Šulc (Eds.). 2007. *Studie z korpusové lingvistiky 2 Kolokace*: 56-93. Praha: NLN.

Čermák, František and Věra Schmiedtová. 2003. The Czech National Corpus Project and Lexicography. Murata, M., S. Yamada and Y. Tono (Eds.). 2003. *Asialex '03 Tokyo Proceedings: Dictionaries and Language Learning: How Can Dictionaries Help Human and Machine Learning?*: 74-80. The Asian Association for Lexicography.

Hanks, Patrick. 2009. Compiling a Monolingual Dictionary for Native Speakers. *Lexikos* 20: 580-598.

Hanks, Patrick. Forthcoming. *Lexical Analysis: Norms and Exploitations.* Cambridge MA: MIT Press.

Hartmann R.R.K. and G. James. 1998. *Dictionary of Lexicography*. London/New York Routledge.

Appendix 1: Lexical Profile of *měkký* in Czech

1-SCHOPNOST a ÚČINEK konkréta fyzický pro agenta (živ): *pod vlivem tlaku n. síly snadný tvarovat, krájet, řezat či skládat, pružný, elastický a neodolný*
a-hmota, materiál, výrobek: *s hladkým povrchem a příjemný na dotek*
b-předmět, výrobek: *zaoblený, ne hranatý*
c-ovoce ap.: *velmi zralý*
d-jídlo: *uvařený, připravený k jídlu*

2-ÚČINEK konkréta/abstrakta fyzický, zvl. akustický, vizuální a hmatový, na receptora (ne/živ): *mající příjemnou vlastnost zahrnující spíše jemný účinek či kontrast než ostrost, vyhraněnost*
a-hlas, zvuk: *tichý a jemný*
b-déšť ap.: *ne silný ani prudký*
c-kontakt, pád, úder: *neprudký*
d-konsonant: *vyslovovaný jako frikativa, třený*

3-ÚČINEK abstrakta psychický na receptora (živ): *sympatický, shovívavý a soucitný, někdy přehnaně*
a-slova, jazyk: *ne příkrý ani rozzlobený, smířlivý*
b-člověk v jednání/projevu: *smířlivý v politice n. nerozhodný, neprůrazný*
c-norma, rozsudek: *nepřísný, nezásadový*

4-ÚČINEK konkréta/abstrakta *jiný* na receptora (živ)
a-nápoj a jiná látka: *působící slabou měrou*
b-trh, měna, zboží: *klesající na hodnotě*
c-voda: *bez minerálů (a nevhodná mj. na holení)*
d-droga: *nenávyková*
e-radiace: *málo pronikavá*
f-pornografie: *spíš náznakově, neexplicitně erotický*

Appendix 2: Sample concordance of the lemma *měkký*, organized semiotically (with annotation)

Concretes

14: Tenkrát jsem spal taky na *slámě*, jenže byla <**měkčí**> . Tahle tlačí a píchá. Chtělo by to posta K
15: jistě víte z teorie i praxe, jsou *dřeva* tvrdá a <**měkká**> . TVRDÁ mají hustá vlákna, a proto se K
18: ;nelidský řev, kolo se přehouplo přes *cosi* <**měkkého**> , a Prokop se probudil. Nahmatal, že K
19: astné, a na to holštýnský řízek právě dost <**měkký**> , aby lahodil patru, s dozlatova opečený K
20: io, s hmyzím *soustem* přesně tak velkým a <**měkkým**> , aby zachutnalo jeho ochmýřené, ro K
21: ulisáci zapomněli pod hradby položit *něco* <**měkkého**> , aby měla na co dopadnout. Výsled K
23: stane. Sádra se nejlépe rozdělává v *nádobě* <**měkké**> , buď speciální gumové misce, která je K
24: nu se večer ochladilo, lehce přimrzalo a v <**měkkém**> , chladném *vzduchu* bylo cítit závan j K
27: ne na Žižkov. *Terén* hřiště U Nisy je zatím <**měkčí**> , do neděle však pravděpodobně zmrz K
29: barevných kovech jen stručně : *MOSAZ* je <**měkká**> , dobře se zpracovává, bývá pěkně žlu K
30: átečníky a pro pokročilé. - Coby softcarver <**měkká**> , dobře ovladatelná, bezproblémová *ly* K
31: my. Nehty a vlasy *Nehty* novorozence jsou <**měkké**> , dosahují konečků prstů, často je i pře K
33: tvrdě, jak doufala. Slunce svítilo za mraky. <**Měkký**> , hedvábný *déšť* padal mezi borovice K
34: šest kilogramů, které naše hlava váží, totiž <**měkké**> , hlavně pak vysoké *podušky* vůbec ne K
35: stavil a člověk mohl pozorovat pohyb jeho <**měkkých**> , jakoby vycpaných *tlapek*, to jak se j K

Abstracts

25: šťuje syntezátory a rozeznívá jimi zejména <**měkkou**> , chrámově varhanní *atmosféru*. Své h A
28: u, pak v za jeho drsnou slupkou objevíme i <**měkké**> , dobré *jádro*, pak vycítíme, že za jeho A
37: osudu nebylo pouhou náhodou, že by jeho <**měkké**> , jemné, nehmotné *jméno* odmítalo sp A
41: e ho zmocňuje *cosi* nevýslovně obrovského, <**měkkého**> , lehkého, průsvitného a přečistého A
47: radace je charakterizována stupnicí : velmi <**měkká**> , **měkká**, měkčí, normální, tvrdší, tvrd A
48: . alternativních scénářů, nabízejících jakýsi <**měkčí**> , mírnější, ohleduplnější nebo " sociálně A
51: Hudák. To, že *ekonomika* byla vlastně příliš <**měkká**> , nakonec musela přiznat i koalice ve A
55: lyrickým pasážím, které tolik vyhovují jeho <**měkkému**> , něžnému a civilnímu *projevu*. I os A
56: d of Paradise ?... brumendem převzali kluci <**měkký**> , něžný *chorus*, jako hučení lesa... hey A
57: kterizována stupnicí : velmi **měkká**, **měkká**, <**měkčí**> , normální, tvrdší, tvrdá, velmi tvrdá A
61: chtěl. Co je tvrdé, vzdorné, to se zlomí. *Co* je <**měkké**> , poddajné, to se ohne, ale nezlomí. C A
72: tři palce od jeho čenichu, a hovořila k němu <**měkkým**> , sípavým *pokuckáváním*, co chvíli A
81: na a jako Varvara dala tušit, že její sametově <**měkký**> , tmavý *mezzosoprán* neztratil nic ze s A
97: paláci. Galerie Velryba, jejíž problematicky " <**měkká**> " *koncepce* zahrnuje kvalitativně nev A
101: i na majitele Objevily se již spekulace, že " <**měkký**> " *postup* ČNB je motivován předvole A

Humans

5: sdržnost. To beru velmi vážně a nemíním být <**měkký**> . Ale na druhé straně se nemíním vy H
10: né s předváděním a přednáškou. Jsem *člověk* <**měkký**> . Pokaždé je mi prodávajícího líto, ž H
17: em si říkal, že letos se na to vykašlu, ale jsem <**měkkej**> . Uvědomujete si, nakolik Lucie ovl H
44: ouchejte, slečno Meg ! Když ste v životě moc <**měkká**> , lidi vás využívají. To si pamatujte ! H
50: době mnohými viděn jako *člověk* zbožný, ale <**měkký**> , muž kompromisu. Arcibiskup Bera H
68: roto, že si myslím, že *lidé* zkrátka jsou takoví. <**Měkcí**> , přizpůsobiví, slabí, a proto chtiví, z H
71: musí být škvíra a ona mi ji ucpává. *Pepinka* je <**měkká**> , sametová, je moje. Cítím její lepka H
88: ávoji... Prokop měl oči plné slz ; cítil se sláb a <**měkký**> , že se až styděl. Před šestou se však H
90: vel a ti jeho kamarádi nebyli tenkrát tak tuze <**měkcí**> ! Kdyby ten Pithart neslyšel trávu rů H
92: ojena. Nejvíc jí vadilo, že manžel byl " takový <**měkký**> ". Otce popisovala jako autoritativ H
135: té). Nový hlavní konstruktér Mišin byl však <**měkký**> a nerozhodný. Projekt L - 1 nedok H
172: být právě tak bezmocný jako ten nesmělý a <**měkký**> *člověk*, jímž opovrhoval. Žena proh H

238: , maličká. Jsme tvrdé jako kámen a zároveň <**měkké**> jako dětská bačkora. Copak já vím, H
246: - AP Tvrdý obchodník z Dallasu je v jádru <**měkký**> KENNEDY BUDE MATKOU V tém H
289: umí plést hebké svetry, v politice rozhodně <**měkká**> není, " soudí znalci, kteří bedlivě sle H
365:ční středisko pro lidská práva označuje jako <**měkké**> skinheady, dohlížely desítky polici H

Animal

69: se krmič opije a nepřijde. Některé *krávy* jsou <**měkké**> , pustí mléko samy, ale většinou m An
75: hnízdu, v němž seděla vrkající *holubice*, celá <**měkká**> , šedá, krásná - nádherný výtvor for An

Metaphors

7: či všem drogám, ani faktickou legalizací *drog* <**měkkých**> . Jenže právě toto " tvrdé jádro " o M
52: a lety byly v centru pozornosti policie *drogy* <**měkké**> , např. marihuana a hašiš, dnes už ve M
85: tomilí. *Opravila* jsem pak v duchu tvrdé *y* na <**měkké**> , uvědomujíc si je všechny tři. Ivana, M
96: lze však sotva očekávat, že by Dánové měli " <**měkčí**> " azylovou *politiku* než zbytek EU, p M
105: extem Bradleyho Strattona posluchačům s " <**měkčíma**> " *ušima* jako by tlumočí stoneovs M
141: pravdu zavřela brána. Za tím *krajem,* který je <**měkký**> a sladký jako tělo, a na čele hlavy M
151: Napsala omýtka s tvrdym *y* a dobili hrad s <**měkkym**> a zapoměl s ie a dokonce ve slově M
182: desetiletí odvážnou cestu uvolnění prodeji <**měkkých**> *drog.* Tento experiment přinesl ús M
183: á dohromady album na podporu legalizace <**měkkých**> *drog.* Účast zatím přislíbili mimo M
199: t na tvrdých drogách, jako je heroin. Přitom <**měkké**> *drogy* jako marihuana, jsou prý na s M
200: provázkem, drátem a podobně. Při vázání <**měkkých**> *dřev* dejte pozor, aby provázek ne M
267: Je třeba, abychom si vzájemně porozuměli. <**Měkká**> křídla evropského Fénixe Marcell v M
376: ního odběratele našich výrobků, které se na <**měkkém**> sovětském trhu nemusely příliš s M

Terms

211: láte ostrou špičku. Sklo podložíte plstí nebo <**měkkým**> dřevem a místo, kde má být díra, T
217: tává než pramínek ušlechtilosti nafilmované <**měkkými**> filtry, domnívá se list a píše o pr T
235: m v jeho testu se vyskytovali Přemyslovci s <**měkkým**> i. Zarážející jsou rovněž gramatick T
256: způsoby ukončení řádku je zásadní rozdíl. <**Měkký**> konec řádku dokáže editor při další T
257: jší akcí, nemůže dojít. Podobně jako tvrdý a <**měkký**> konec řádky, existuje i tvrdý a **měk** T
258: dý a **měkký** konec řádky, existuje i tvrdý a <**měkký**> konec stránky. **Měkký** konec stránk T
259: ky, existuje i tvrdý a **měkký** konec stránky. <**Měkký**> konec stránky vytváří editor podle z T
261: RÁK o šíři čelistí 60 až 80 mm s vložkami z <**měkkého**> kovu, protože bez něho nemohou T
262: lých hmot dát na čelisti vložky ve tvaru L z <**měkkého**> kovu (olova, hliníku), aby se pře T
263: jeden jemný na tvrdé kovy, jeden hrubší na <**měkké**> kovy, tvrdé umělé hmoty a dřevo v T

Idioms

377: u Manuelou nijak zvlášť nestál, " mám totiž <**měkké**> srdce, padre. " " **Měkké** srdce je do IF
378: doufat, že jste udělala dobře, ale máte příliš <**měkké**> srdce ! " Vtom lázeňské přivedly tři IF
379: jí). Ale na druhé straně mi teď došlo, že má <**měkké**> srdce (vždyť zatajila svou totožnost IF
380: nestál, " mám totiž **měkké** srdce, padre. " " <**Měkké**> srdce je dobrá věc, synu, ale v přípa IF
381: ana za roli ve snímku Frajer Luke, mužem s <**měkkým**> srdcem. Svědčí o tom i jeho vzta IF