TOWARDS A NEW TYPE OF MORPHEMIC ANALYSIS

Eva Koktová
9. května 1576
39001 Tábor, Czechoslovakia

ABSTRACT

The present paper provides a report on a new system of an automated morphemic analysis of technical texts in Czech as a highly inflectional language, which is being prepared by the linguistic team of the Faculty of Mathematics and Physics in Prague, within the project of man-machine communication without a pre-arranged data base (TIBAQ). The kind of morphemic analysis presented here is based on a retrograde (right-to-left) analysis of words by means of morphemically unambiguous or irresolvably ambiguous word-ends, which do not coincide with the etymological word-endings but correspond to the structure of the accidental cases of morphemic ambiguity in an inflectional language (word-endings being accountable for in a certain way by word-ends). The algorithm of analysis can thus dispense with any dictionary (of morphemic irregularities and exceptions), economically accounting especially for productive word-endings. The word-ends of the analysis are assigned several kinds of morphemic information, concerning morphemic categories and lemmatization. The analysis is based on the absolute frequency of word-ends in technical texts and is able to interact with the semantic analysis.

1. INTRODUCTION

The present paper provides a report on a new system of an automated morphemic analysis of technical texts in Czech, which is being prepared by the linguistic team of the Faculty of Mathematics and Physics in Prague. The morphemic analysis of Czech, which is a highly inflectional language, constitutes the starting point for any kind of automated processing of language, ranging from automatic information retrieval to natural language understanding.

There is a previous project of morphemic analysis of Czech described in (Weisheitelová, Králíková and Sgall, 1982), which is based on an analysis of etymological word-stems and word-endings (suffixes). The present system, on the other hand, is based on a retrograde (right-to-left) analysis of words, which makes it possible to dispense both with the dictionary of stems and the dictionary of endings; it was partly inspired by the system MOSAIC (Kirschner, 1982) (intended first of all for automatic indexing of technical texts), which is also based on a kind of retrograde analysis: namely, on singling out the four rightmost symbols of the word-forms of autosemantic words, which are then matched against a list of word-endings. This kind of analysis, however, cannot avoid the danger of ambiguity, which is prevented by a number of ad-hoc restrictions, for example reducing the universe of discourse.

The present system of morphemic analysis differs from the previous ones in several essential respects:

(i) The algorithm of the present type of morphemic analysis can be viewed as a structured list of morphemically unambiguous or irresolvably ambiguous word-ends of Czech words (which may be accidentally identical with full word-forms) including information concerning their morphemic categories and lemmatization. We believe that this principle can be considered as adequate for the morphemic analysis of any inflectional language.

(ii) In the present system, it is also easier to carry out lemmatization: there are only several tens of simple and highly general lemmatization rules appended to the morphemic information accompanying every word-end in the algorithm.

(iii) In the present system, the burden of the analysis lies entirely on the algorithm. There is no need of any dictionary in which etymological irregularities would be listed.

(iv) The algorithm is based on the absolute frequency of word-ends in technical texts. It consists of two parts; the first of them involves about two hundred word-ends by means of which it is possible to resolve about fifty percent of a technical text.

(v) By means of the algorithm it is possible to analyze an unlimited number

of new (newly coined) words with productive etymological word-endings. Thus, both the user and the linguist are relieved of the work which must be usually done when a new lexical item is being incorporated into a system of morphemic analysis of an inflectional language.

(vi) The algorithm is going to be implemented in PL/1 within a system of natural language understanding, namely the project of man-machine communication called TIBAQ (Text-and-Inference Based Answering of Questions, cf. (Hajičová and Sgall, 1981)) with no pre-arranged data base and with the capacity of self-enriching by information drawn from the text; the project is based on the linguistic theory of the Functional Generative Description.

(vii) Underlying the algorithm is a large amount of empirical work; it analyzes several tens of thousands of (autosemantic and synsemantic) words (drawn from a retrograde dictionary of Czech, cf. (Slavíčková, 1975)), including the word-forms of inflected words. The choice of the autosemantic lexical units to be analyzed was carried out with respect to technical texts concerning microelectronics.

2. THE PHILOSOPHY OF THE SYSTEM

The major novelty of the present approach consists in the conception of (morphemically unambiguous or irresolvably ambiguous) word-ends, which do not correspond to the (etymological) word-inflection and word-formation endings but to the cases of accidental morphemic ambiguity in an inflectional language, every word-ending being accountable for by at least one word-end (piece of output information). On the other hand, every word-end corresponds to (stands for) at least one lexical word, and due to the cases of morphemic ambiguity, it represents at least one word-form. A word-end is usually equivalent to a part of a word-form, but accidentally it may be equivalent to a full word-form.

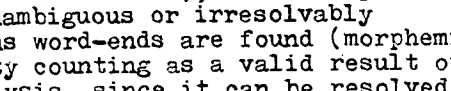
The algorithm of analysis, embodying a conception of procedural morphemics, can be viewed as a structured list of word-ends arranged in a branching structure consisting of yes-no answers to queries, with corresponding sequences (strings) of symbols of increasing length, which is due to the retrograde adding of symbols (we use 40 letters of the Czech alphabet, including the ones with diacritics), until morphemically unambiguous or irresolvably ambiguous word-ends are found (morphemic ambiguity counting as a valid result of the analysis, since it can be resolved, in most cases, by means of the syntactic analysis). The word-ends are assigned the kinds of information as described in section 3.

In the present system of morphemic analysis, there is no place for the notion of (etymological) irregularity, all word-ends being equally "regular"; the differences between them can be accounted for e.g. in terms of their length or of their positions on the scale of absolute frequency (cf. section 5). It may even be the case that an etymologically highly irregular word-form can be analyzed by a relatively small number of symbols (of its word-end), and the other way round.

In the horizontal progress of the algorithm (which corresponds to the answer yes - a new symbol is added) the output information concerns a single word-end, while in the vertical progress (corresponding to the answer no - different symbols than the one(s) in question are added) it usually concerns more than one word-end. These word-ends can be labelled as complementary word-ends with respect to the horizontal word-end(s) in question; they consist of the same sequence of symbols as the correlated horizontal word-ends with the exception of their respective leftmost symbols, which belong to the complementary set of symbols of the alphabet with respect to the leftmost symbol(s) of the horizontal word-end(s), according to the combinatorics of letters in existing Czech words (for example, the complementary word-ends to the horizontal word-ends /měr, úměr, ýměr are only four: áměr, íměr, oměr, ůměr (the symbol / stands for the end of the word, i.e. indicates a word-end in the form of a full word-form)). Throughout the algorithm, the notation concerning the complementary word-ends is abbreviated in that in their place only their common output information is written (cf. the three occurrences of A in Figure 1 below).

The conception just discussed can be illustrated by a chunk of the algorithm accounting for the frequent word-inflection ending ý (which is an adjectival word-ending, ambiguous among nominative and accusative singular masculine-inanimate, and nominative singular masculine-animate, thus representing the adjectival "normal form"), which clashes only with /rý (adverb), being accounted for by the three occurrences of the output information A (standing for the morphemic information in question) in Figure 1.

Figure 1. A chunk of the algorithm.



The three occurrences of A in Figure 1 can be indicated, for the sake of clarity, as A_1, A_2 and A_3: A_1 (corresponding to the

horizontal string rý) accounting for those Czech adjectives (in the given form) whose penultimate symbol is different from r (such as velký (big)), A_2 (corresponding to the horizontal string prý) accounting for those Czech adjectives (in the given form) whose second symbol from the right is r and whose third symbol from the right is different from p (such as dobrý (good)), and A_3 (corresponding to the horizontal word-end /prý) accounting for those Czech adjectives (in the given form) whose third and second symbols from the right are pr, respectively, and whose fourth symbol from the right is different from /, i.e. which are longer than three symbols (in Czech, there is only one such adjective, namely kyprý (loose, plump)). On the whole, A_1, A_2 and A_3 account for all Czech adjectives (in the given form).

3. KINDS OF INFORMATION

The word-ends (i.e. the horizontal word-ends and the complementary word-ends with respect to the given horizontal word-ends) are assigned the following kinds of information.

A. Morphemic information.

(i) The information concerning part-of-speech categories includes the distinction between Nouns, Verbs (these kinds of information are further subcategorized), Adjectives (A), Adverbs (B), Prepositions (C), Conjunctions (D) and Pronouns (Zj) (there are distinguished three kinds of pronouns, namely those which function as nouns, those which function as adjectives, and those which function both ways).

(ii) The information concerning grammatical categories includes the following distinctions (with respect to the part-of-speech categories).

(a) Declension.

(aa) Case (six cases, indicated as 1, 2, 3, 4, 6 and 7) is distinguished not only with nouns, but due to grammatical agreement, also with adjectives and pronouns.

(bb) Number (singular and plural, indicated as sg and pl, respectively) is distinguished with nouns, and due to grammatical agreement, also with adjectives, pronouns and verbs.

(cc) Gender (combined with animateness) is distinguished with nouns, and due to grammatical agreement, partly also with adjectives, pronouns and verbs (with verbs, for example, in the past and passive participles plural). With nouns, four genders are distinguished: masculine-inanimate (N), masculine-animate (Ž), feminine (F), and neuter (S). The category of animateness is involved rather with masculine than with feminine and neuter nouns because with plural masculine nouns the difference in animateness is present, due to grammatical agreement, also with verbs and adjectives in the above mentioned way, and because in technical texts substantially more masculine-animate than feminine-animate nouns are found.

(b) Conjugation.

With verbs, there is distinguished person (three persons, with the exception stated in section 4), number (cf. (bb) above), tense (present, past and future), mood (indicative and imperative), and voice (active and passive). As concerns notation, usually several kinds of information are collapsed in a single abbreviation, cf. K standing for the third person singular active indicative present.

There is no need of information concerning the inflectional types of nouns, adjectives and verbs; for example the word-ends corresponding to the class of nouns represented by the word-forms katodami (by cathodes) and vlastnostmi (by properties) (both 7 pl) are assigned the same morphemic information, though the word-forms in question belong to etymologically quite different types of inflection of (feminine) nouns (cf. the difference between the word-inflection endings, ami and mi, respectively).

B. Lemmatization information.

Lemmatization, i.e. convering an inflected word-form into the normal form (i.e. 1 sg with nouns, 1 sg masculine with adjectives and pronouns, and the infinitive form with verbs) has a specific purpose, being connected with those applications of morphemic analysis which concern the terminological elements of technical texts (such as automatic indexing).

In the present system, lemmatization is carried out by a retrograde erasing of a certain number of symbols (possibly zero) and by adding a number of specific symbols (possibly zero) to what has been left after the erasing; in lemmatization (unlike in the rest of the algorithm) we work with diacritic marks as specific symbols. In this way, lemmatization can be accounted for by means of several tens of simple and highly general rules, cutting across the inflectional endings and also across the inflectional types of different part-of-speech categories. It should be pointed out that lemmatization concerns rather the concrete words (word-forms) found in a text than the word-ends themselves: though the majority of the lemmatization rules operate on word-ends (concerning usually only a part of a word-end, which is close to a word-

-ending, cf. the symbol y in the word-end tody, corresponding to the word-form katody), in exceptional cases, for example where the stem of a word is affected by an alternation, the erasing may reach to the left of the concrete word, i.e. behind the word-end; cf. the word-end ste (consisting of three symbols), which, with some simplifications, unambiguously indicates a verb (K), but which is not sufficient for the lemmatization of such verb-forms as roste (grows) to their infinitives (růst (to grow)), where four rightmost symbols of the concrete word should be considered.

The rules of lemmatization have generally the form [X; abc...], where X stands for the number of the symbols to be erased, and abc..., for the specific symbols to be added. In the algorithm, the rules are usually referred to by numbers, and listed in an appendix. Thus, for example, Rule 2 ([1; a]) converts katody (cathodes; F 2 sg & 1 & 4 pl) into katoda (cathode; F 1 sg) by erasing one symbol (namely y) and by adding one symbol (namely a). (& stands for the relation of ambiguity).

Every lemmatization rule has at least one application to various types of morphemic categories concerning not only different distinctions within a single part-of-speech category (typically, different genders with nouns) but also different part-of-speech categories (for example, a single lemmatization rule can be applied to nouns, adjectives, and verbs); this means that a lemmatization rule may concern, in any of the part-of-speech categories in question, more than one word-ending (e.g. of different gender), and these word-endings may be in turn ambiguous between various case-and-number assignments.

This can be illustrated by Rule 6 and Rule 8. Rule 6 ([1; ∅] - erase one symbol, add nothing) cuts across nouns, adjectives, and verbs, converting e.g. spoje (communications) to spoj (communication); mladým (to young ..., to young ...) to mladý (young), and vysáty (sucked up) to vysát (to suck up). Rule 8 ([2; ∅] - erase two symbols, add nothing) has 9 applications (to all genders of nouns and to adjectives) and corresponds, on the whole, to 16 word-endings, out of which two are two-ways ambiguous as concerns case and number. The 16 word-endings are illustrated by the word-forms in Figure 2 (where obvod = circuit, odborník = expert, katoda = cathode, vlastnost = property, relace = relation, stavení = building, mladý = young, and původní = original).

Figure 2. Lemmatization.

N: obvodě (6 sg); obvodem (7 sg); obvodů (2 pl)

Ž: odborníkem (7 sg); odborníků (2 pl)

F: katodám, vlastnostem (3 pl); katodami, vlastnostmi, relacemi (7 pl)

S: staveních (6 pl); staveními (7 pl)

A: mladých, původních (2 & 6 pl); mladými, původními (7 pl)

In the above survey, the words which are assigned common information (e.g. katodami, vlastnostmi, relacemi) belong to etymologically different types of inflection, which, however, need not be distinguished here: though the lemmatization rules can be arranged in a scale according to their complexity or range of application, the present method of lemmatization covers both simple (regular) and complicated (irregular) types of word-inflection and word-formation in an equally economic manner.

C. Semantic information.

The semantic analysis by means of the retrograde morphemic analysis is a yet unfinished, but presumably smoothly feasible task, which will be based on the account of productive word-endings by means of word-ends.

The considerations concerning the semantic analysis should start from establishing a set of semantic categories (classes) of nouns and possibly also adjectives which are considered to be relevant for the analysis of technical texts. In addition to the consideration of productive word-endings, there can be also introduced into the algorithm such word-ends which account for semantically relevant but only restrictedly productive word-formation endings (such as metr (meter)), if such word-ends have been "hidden" in the complementary word-ends of the algorithm (for example, it may happen that a productive word-ending coinciding with a single word-end (such as tko, cf. below) is "hidden" in this way).

In establishing the set of semantic categories, we can draw from (Buráňová, 1980) and (Kirschner, 1983), proposing that there should be introduced for example the category of Instrument (Tool) (as expressed by the productive word-endings dlo, tko, ač, ič, čka, ér, or, and by the restrictedly productive word-endings metr, graf, fon, and skop), Action (Process) (ace, kce, aní, ání, ení, ění, áž and za), Property (ost, ita and ance), etc.

The information concerning semantic

analysis can be rendered by indicating certain pieces of output information as semantically relevant (with respect to the classification of semantic categories), but presumably it will be even possible to state this kind of information essentially only in an appendix to the algorithm. Such an appendix should consist of the specification that every word-end (this concerns also complementary word-ends) whose rightmost symbols coincide with the word-ending in question (because a word-end is usually longer than, or identical to, the word-ending which is accounted for by it) and which is assigned certain morphemic information (concerning usually gender) corresponds to the semantic category in question; cf. all word-ends whose three rightmost symbols are ací and which are assigned the output information F 7 sg § 2 pl (such as lací, which is "hidden" in the complementary word-ends) correspond to the semantic category of nouns of action (in this case, ací is correlated to the normal form with ace, which is the Czech equivalent of the English ation). Possible exceptions to the semantic information concerning the word-ends which account for the word-endings in question should be indicated directly in the algorithm (e.g. by superscripts in the output information); for example, the above-mentioned nominal word-ending ací (which systematically clashes with the adjectival word-ending ací N § F § S 1, 4 sg § Ž 1 sg § F 2, 3, 6, 7 sg § N § Ž § F § S 1, 4 pl, and thus is accounted for by about 30 pieces of output information) has about five semantic exceptions to it (such as nadací (nadace = grant, support - neither action nor result of action)), for which there should be established special word-ends in the algorithm, with the indication, in the output information, of their semantic exceptionality (with respect to the other word-ends whose rightmost symbols are ací and which are assigned the output information in question), i.e. of their non-membership in the class of nouns of action.

4. AMBIGUITY

This section brings information concerning (i) cases of morphemic distinctions not included in the algorithm; (ii) genuine irresolvable cases, and (iii) cases of morphemically irresolvable ambiguity.

(i) Cases of morphemic distinctions not included in the algorithm. We prefer not to include in the algorithm of analysis (with possible exceptions) morphemic distinctions concerning those word-inflection endings which occur in technical texts only rarely or not at all, particularly the following distinctions:

(a) Verbs: 1 sg indicative present (such as předpokládám (I suppose)); 2 sg indicative present (such as předpokládáš (you suppose)); 2 sg imperative (such as vyber (choose)); transgressive forms (such as předpokládaje, předpokládajíc, předpokládajíce (supposing)), and 1 and 2 pl imperative are assigned only the morphemic but not the lemmatization information because these forms are supposed not to be semantically relevant.

(b) Nouns: 5 sg and pl (such as odborníku! (expert!)).

(c) Adjectives: masculine-animate pl (such as vysocí (tall)).

(ii) Genuine irresolvable cases. By the present kind of analysis, there practically cannot be resolved, in spite of their regular inflection, geographical and personal proper names, their multitude preventing the linguist from empirically establishing their (unambiguous or ambiguous) word-ends. This can be partly overcome by introducing into the analysis the recognition of capital letters and/or by establishing a "right set" of proper names to be analyzed (which seems to be an easier task with geographical names, cf. Evropa (Europe), Praha (Prague), etc.). On this solution, for example, the accusative form of Praha (F), namely Prahu, would yield a case of morphemically irresolvable ambiguity with the locative form of práh (N; threshold), namely prahu. Also certain frequent personal names can be treated in this way (cf. Schottkyho dioda (the diode of Schottky)).

(iii) Cases of morphemically irresolvable ambiguity. The cases of this kind of ambiguity concern all of the morphemic categories as well as lemmatization, occurring singly or as combined in various ways. In what follows, the relevant cases of ambiguity are indicated by §, and the other cases of ambiguity are inducated by commas or semicolons.

(a) Ambiguity concerning only part-of-speech category; cf. the ambiguity of the word-ends corresponding to non-inflected words, such as the ambiguity of the word-end tř between advert and preposition (D § C), tř standing for several words including e.g. vevnitř (inside) or zevnitř (from inside).

(b) Ambiguity concerning part-of-speech category in combination with other kinds of ambiguity; cf. the ambiguity of the word-ends corresponding to inflected words, such as the ambiguity of the word-end /růst between noun and verb (N 1, 4 sg § Infinitive: growth § to grow), or the ambiguity of the word-end /rovná between adjective and verb (A F 1 sg; S 1, 4 pl § K: direct § straightens).

(c) Ambiguity concerning only gender, cf. the ambiguity in gender concerning word-inflection endings with adjectives, such as the ambiguity of the word-ends (coinciding, with one exception, with word-inflection endings) ých (2, 6 pl) and ými (7 pl), which are ambiguous among all genders (N § Ž § F § S).

(d) Ambiguity concerning gender in combination with other kinds of ambiguity:

(aa) Ambiguity concerning gender in combination with case and number, cf. the word-end /set, which is ambiguous between masculine-inanimate and neuter noun (N 1, 4 sg § S 2 pl: set § of hundreds).

(bb) Surface-syntax ambiguity concerning gender in combination with underlying ambiguity concerning case and number, cf. the word-end /řádky (lines), which is ambiguous between masculine-inanimate and feminine noun (N 1, 4, 7 sg § F 2 sg; 1, 4 pl). This ambiguity in gender, however, is not present on the underlying level of Czech, where only a single lexical item (masculine-inanimate noun) is hypothesized to occur, as corresponding to the two surface normal forms (i.e. masculine-inanimate and feminine), the two surface genders accidentally yielding ambiguity in the word-end (word-form) /řádky.

(cc) Ambiguity concerning gender in combination with animateness (and case), cf. the word-end /člen (member), which is ambiguous between masculine-inanimate and masculine-animate noun (N 1, 4 sg § Ž 1 sg). (In the majority of the other cases of the inflection of masculine nouns, the ambiguity in animateness is not accompanied by the case ambiguity.)

(e) Ambiguity concerning only case (and number), not accompanied by any other kinds of ambiguity, cf. the word-end tody (F 2 sg § 1 § 4 pl).

(f) Systematic ambiguity concerning the distinction between geographical names and possessive adjectives derived from lexically corresponding personal names, cf. the word-end /Benešova (N 2 sg § A N 2 sg; F 1 sg; S 1, 4 pl: of Benešov § of Beneš's).

(g) Ambiguity concerning lemmatization, cf. the word-end vyváží (K), corresponding to a single word-form vyváží, between lemmatization rules [1; t] and [2; et], corresponding to the infinitives vyvážit (to balance) and vyvážet (to export), respectively. Cf. also the surface-syntax ambiguity in lemmatization with the word-end /řádky (cf. (bb) above), which is surface-syntax ambiguous in gender (N: řádek § F: řádka).

The present treatment of ambiguity is characteristic of the procedural conception of morphemics in that the method of accounting for every etymological word-ending by means of at least one word-end (piece of output information) removes from the analysis the systematic ambiguity as well as morphemic irregularities (exceptions) concerning etymological word-inflection and word-formation endings, which have been usually treated by means of various restrictions and other ad-hoc means. Every case of the systematic etymological ambiguity is accountable for by several tens or even hundreds of pieces of output information (cf. the systematic ambiguity of the word-formation ending ací as mentioned in section 3, or that of the word-inflection ending y among masculine-inanimate, masculine-animate and feminine nouns with additional morphemically irresolvable ambiguity concerning case and number: N 1, 4 7 pl § Ž 4, 7 pl § F 2 sg; 1, 4 pl); on the other hand, exceptions to word-endings (in the form of word-ends with different output information) are accountable for by several pieces of output information (cf. the word-inflection ending ý as mentioned in section 2, which is accountable for by three pieces of output information, representing one exception, or the word-formation ending ení as mentioned in section 5, which is accountable for by five pieces of output information, representing six exceptions).

After resolving the cases of the systematic etymological ambiguity and of irregularity, it is possible to list the remaining (about one hundred) cases of morphemically irresolvable ambiguity (with the exception of the case-number ambiguity accompanying gender ambiguity); such a list can be compared to the list by (Panevová, 1981) involving ambiguous word-forms in Czech. Panevová's list, not being lexically restricted with respect to specific applications, includes also proper names, words not occurring in technical texts and forms not analyzed by the present algorithm (such as singular imperative with verbs), but on the other hand, it consists only of full word-forms, thus intersecting with the present list, where first of all ambiguous word-ends in the form of parts of words are involved.

5. QUANTITATIVE ASPECTS

The present conception of the algorithm of morphemic analysis is based on the absolute frequency of word-ends in technical texts. In the ideal case, the word-ends should be arranged with respect to the frequency of their last (rightmost), last-but-one, etc., symbols - a task which itself would require the aid of a computer; for the time being, we must

work with an approximation, which makes it necessary to divide the algorithm into two parts according to the assumption that the first two hundred word-ends on the scale of absolute frequency, arranged according to a statistical examination concerning the whole word-ends, could resolve about fifty percent of the words of a technical text, while the other word-ends of the algorithm (pieces of output information), arranged according to the frequency of their last symbols, should resolve the remaining portion of a technical text. We assume that out of the about twenty thousand pieces of output information of the broadly conceived preliminary version of the algorithm, only several thousands will be sufficient to cover the words which may occur in a standard technical text (this will lead to a substantial reduction of the preliminary version of the algorithm).

The words included into the analysis fall into four major semantic hyper--categories (not used in the semantic analysis): (i) words with the most general semantics (including the forms of categorial verbs, such as být (to be), prepositions, such as v (in), etc.); (ii) general terms typical of technical texts (such as metoda (method), systém (system), etc.); (iii) words specific to the given technical domain, e.g. microelectronics (such as katoda (cathode), obvod (circuit), etc.), and (iv) words typical of other (possibly affiliated) domains (such as cihla (brick), střecha (roof), etc.).

The conception of the most frequent two hundred word-ends (which are arranged in a special algorithm) can be illustrated by a list involving ten most frequent word-ends; in Czech technical texts, they belong to the first hyper-category. These word-ends are of three kinds: (i) word-ends in the form of parts of word-forms (which may accidentally coincide with etymological word-endings, such as /ch or ého); (ii) word-ends in the form of full word-forms (such as /se or /ie), and (iii) word-ends in the form of parts of word-forms resolvable with minor exceptions (such as (ý) or (ení)); such word-ends are indicated by encircling. In addition to this, there can be distinguished morphemically unambiguous word-ends (cf. /na, /a, /v, uje) vs. morphemically ambiguous word--ends (cf. /ch, /se, /ie, (ení), ého, (ý)). In the list in Figure 3, all cases of ambiguity (including the ambiguity in case and number) are indicated by §; with /ie, for the sake of clarity, the morphemic information is given directly by means of English equivalents.

Figure 3. Frequent word-ends.

1. /ch -- A N § Ž § S 2 § 6 pl
2. /se -- Zj (reflexive) § C (with)
3. /ie -- is § them (4 pl) § it (4 sg)
4. (ení) -- S 1 § 2 § 3 § 4 § 6 sg § 1 § 2 § 4 pl
5. /na -- C (on, for)
6. /a -- D (and)
7. /v -- C (in)
8. ého -- A N § Ž § S 2 sg § Ž 4 sg
9. uje -- K
10. (ý) -- A N § Ž 1 § 4 sg § N 4 sg

6. CONCLUSION

We have described a not yet implemented but promising system of a right-to-left morphemic analysis intended for technical texts in Czech and based on a conception of morphemically unambiguous or irresolvably ambiguous word-ends as embodying the cases of morphemic ambiguity in an inflectional language. The present system seems to be more economic than the previous systems (which are fully or partly based on the conception of etymological word-endings (and word-stems) or on the conception of word-ends as consisting of a fixed, apriori established number of symbols) in that it can dispense with any dictionary as well as with the notion of morphemic irregularity; moreover, it is capable of an interaction with the other levels of analysis, as well as of various adjustments.

The advantages of the present system vis-à-vis the previous systems can be summarized as follows.

(i) Due to the fact that every set of complementary word-ends (with respect to the given horizontal word-end(s)) is assigned a common piece of output information, and also to the fact that even a single word-end often corresponds to several words (lexical units) and/or to several word-forms, the number of the pieces of output information necessary for resolving a standard technical text is presumably considerably lower than the number of the word-forms (of both inflected and uninflected words) occurring in such a text.

(ii) The present system is able to account for the word-forms of new (newly coined) words with productive word--endings automatically, without considering their stems.

(iii) The account of productive word--endings also enables to account for semantically relevant word-endings by indicating the semantically relevant pieces of output information.

REFERENCES

1. Buráňová Eva. 1980. Ob odnoj vozmožnosti semantičeskoj klassifikaciji suščestvitel'nych (On one possibility of semantic classification of nouns). Prague Bulletin of Mathematical Linguistics 34, 33-44.

2. Hajičová Eva and Sgall Petr. 1981. Towards Automatic Understanding of Technical Texts. Prague Bulletin of Mathematical Linguistics 36, 5-24.

3. Kirschner Zdeněk. 1982. MOSAIC - A Method of Automatic Extraction of Technical Terms in TExts. Prague Bulletin of Mathematical Linguistics 37, 5-28.

4. ----------. 1982. On a device in dictionary operation in machine translation. COLING 82 - Proceeding of the Ninth International Conference in Computational Linguistics. North Holland - Academia.

5. Konečná D. and Hronek J. 1960. Morfologická analýza podle posledního písmene (Morphological analysis according to the last letter). Acta Universitatis Carolinae: Slavica Pragensia 2. Praha.

6. Panevová Jarmila. 1981. Lexical Input Data for Experiments with Czech. Explizite Beschreibung der Sprache und automatische Textarbeitung VI. Praha: Faculty of Mathematics and Physics.

7. ---------- and Sgall Petr. 1979. Toward an Automatic Parser for Czech. International Review of Slavic Linguistics 4/3, 433-445.

8. Sgall Petr. 1960. Soustava pádových morfémů v češtině (The system of case endings in Czech). Acta Universitatis Carolinae: Slavica Pragensia 2.

9. Slavíčková Eva. 1975. Retrográdní morfematický slovník češtiny (A retrograde morphematic dictionary of Czech). Praha: Academia.

10. Weicheitelová Jana. 1981. Automatic Analysis of Czech Morphemics. Prague Studies in Mathematical Linguistics 7, 225-236.

11. ----------, Králíková Květa and Sgall Petr. 1982. Morphemic Analysis of Czech. Explizite Beschreibung der Sprache und automatische Textarbeitung VII. Praha: Faculty of Mathematics and Physics.