Charles University Prague
Faculty of Mathematics and Physics

Jiří Hana

Two-level morphology of Esperanto

Prague, August 1998

Charles University Prague
Faculty of Mathematics and Physics

Jiří Hana

Two-level morphology of Esperanto

Master thesis

Department: Computer Science
Thesis supervisor: RNDr. Jan Hajič, Ph.D.

Prague, August 1998

Acknowledgements

I would like to thank to Doc. Dr. Petr Chrdle, CSc. the owner of a publishing house KAVA-PECH, Dobřichovice, Czechia, who donated me Esperanto texts for tests of my system.

I am indebted to my supervisor RNDr Jan Hajič, Ph.D. for his counsels and comments. I am also very grateful to Dr. Hana Skoumalová and ing. Alexander Rosen, the discussions with them were very useful for me.

I declare that I wrote the thesis independently and that I used only cited resources.

Prague, August 11, 1998

Table of contents

1 INTRODUCTION 13

1.1 About Esperanto 13

2 ESPERANTO GRAMMAR 15

2.1 Writing and pronunciation 15
2.2 Case and number 15
2.3 Article 15
2.4 Noun 16
2.5 Adjectives 16
2.6 Pronouns 17
2.7 Correlatives 18
2.8 Numerals 20
2.9 Verb 24
2.10 Adverbs 27
2.11 Prepositions 28
2.12 Conjunctions 29
2.13 Interjections 29
2.14 Short syntax overview 29

3 WORD BUILDING 33

3.1 Composites 34
3.2 Affixes 35
3.3 The rest 43

4 IMPLEMENTATION 47

4.1 Two-level morphology 47
4.2 General approach 48
4.3 Inflection 49
4.4 Verb 50
4.5 Roots 53
4.6 Category prohibiting rules 55
4.7 Personal pronouns 56
4.8 Correlatives 56
4.9 Numbers 58
4.10 Countries 60
4.11 The rest 62

5 CONCLUSION 65

RESOURCES 67

APPENDIX A AUXILIARY PROGRAMS 69

Appendix A.1 Conversion to corpus 69
Appendix A.2 Filtering result of analysis 70
Appendix A.3 Conversion of the PIV 70

APPENDIX B OUTPUT OF THE ANALYSIS 73

Appendix A.1 Sample of the morphological analysis 73
Appendix B.2 Words with more than one analysis 77
Appendix B.3 Unanalyzed words 78

APPENDIX C TWO-LEVEL RULES 80

Detailed table of contents

1 INTRODUCTION 13

1.1 ABOUT ESPERANTO 13

2 ESPERANTO GRAMMAR 15

2.1 WRITING AND PRONUNCIATION 15
2.2 CASE AND NUMBER 15
2.3 ARTICLE 15
2.4 NOUN 16
2.4.1 Proper names 16
2.4.1.1 Declination of proper names 16
2.4.1.2 Capitalization 16
2.5 ADJECTIVES 16
2.5.1 Form 16
2.5.2 Comparison 17
2.6 PRONOUNS 17
2.6.1 Personal pronouns 17
2.6.2 Possessive pronouns 18
2.7 CORRELATIVES 18
2.7.1.1 System of correlatives 18
2.7.1.2 Declination of correlatives 19
2.7.1.3 Using parts of correlatives in word building 19
2.7.1.4 Using correlatives in word building 20
2.8 NUMERALS 20
2.8.1 Cardinal numerals 20
2.8.2 Non cardinal numerals 22
2.8.2.1 Ordinal numerals 22
2.8.2.2 Adverbial numerals 22
2.8.2.3 Names of numbers 23
2.8.2.4 Multiplication numerals 23
2.8.2.5 Collectives 23
2.8.2.6 Fractions 23
2.8.2.7 Distribution 23
2.8.2.8 Other topics 24
2.9 VERB 24
2.9.1 Infinitive 24
2.9.2 Vowels of tense 24
2.9.3 Indicative 24
2.9.4 Conditional 24
2.9.5 Imperative 25
2.9.6 Participles, Gerunds, Verbal nouns 25
2.9.6.1 Participles 25
2.9.6.2 Gerunds 25
2.9.6.3 Verbal nouns 25
2.9.6.4 Verbalized participles 26
2.9.7 Complex verbal forms 26
2.9.7.1 Imperfect 26
2.9.7.2 Perfect 26
2.9.7.3 Predicative 26
2.9.7.4 Infinitive complex forms 27
2.9.7.5 Conditional and Imperative active complex forms 27
2.9.7.6 Passive voice 27
2.10 ADVERBS 27
2.10.1 Form 27
2.10.2 Comparison 28
2.10.3 Inflection 28
2.11 PREPOSITIONS 28
2.11.1 Types of prepositive 29
2.12 CONJUNCTIONS 29
2.13 INTERJECTIONS 29

2.14 SHORT SYNTAX OVERVIEW 29
2.14.1 Accusative *29*
2.14.2 Agreement *30*
2.14.3 Word order *30*
2.14.4 Question *31*
2.14.5 Negation *31*

3 WORD BUILDING 33

3.1 COMPOSITES 34
3.1.1 Determination *34*
3.1.2 Coordination *35*
3.2 AFFIXES 35
3.2.1 True suffixes *36*
3.2.1.1 Aĉ 36
3.2.1.2 Eg 36
3.2.1.3 Et 36
3.2.1.4 Um 36
3.2.2 Suffixoids *37*
3.2.2.1 Igi 37
3.2.2.2 Iĝi 37
3.2.2.3 Ado 37
3.2.2.4 Eco 37
3.2.2.5 Ĉj, Nj 37
3.2.2.6 Other suffixoids 38
3.2.3 Prefixes *40*
3.2.3.1 Bo 40
3.2.3.2 Ge 41
3.2.3.3 Mal 41
3.2.3.4 Pra 41
3.2.3.5 Other prefixes 42
3.2.4 Unofficial affixes *42*
3.2.5 Pseudoaffixes *43*
3.3 THE REST 43
3.3.1 Inserted o *43*
3.3.2 Hyphen *43*
3.3.3 Sciences *44*
3.3.4 Names of countries *44*
3.3.5 Abbreviations *44*

4 IMPLEMENTATION 47

4.1 TWO-LEVEL MORPHOLOGY 47
4.2 GENERAL APPROACH 48
4.2.1 Why not generation *49*
4.2.2 Conventions used in the following text *49*
4.3 INFLECTION 49
4.4 VERB 50
4.5 ROOTS 53
4.5.1 Inserted o *54*
4.5.2 Prefix bo *54*
4.5.3 Prefix pra *55*
4.6 CATEGORY PROHIBITING RULES 55
4.7 PERSONAL PRONOUNS 56
4.8 CORRELATIVES 56
4.9 NUMBERS 58
4.10 COUNTRIES 60
4.11 THE REST 62
4.11.1 Sciences *62*
4.11.2 Coordinative composites *62*
4.11.3 Prepositions *63*
4.11.4 Primitive words *63*
4.11.5 Suffixes ĉj/nj *63*
4.11.6 Units *63*

4.11.7 Replacing $\hat{h}$ after r by k *64*

5 CONCLUSION **65**

RESOURCES **67**

APPENDIX A AUXILIARY PROGRAMS **69**

APPENDIX A.1 CONVERSION TO CORPUS 69
APPENDIX A.2 FILTERING RESULT OF ANALYSIS 70
APPENDIX A.3 CONVERSION OF THE PIV 70

APPENDIX B OUTPUT OF THE ANALYSIS **73**

APPENDIX B.1 SAMPLE OF THE MORPHOLOGICAL ANALYSIS 73
APPENDIX B.2 WORDS WITH MORE THAN ONE ANALYSIS 77
APPENDIX B.3 UNANALYZED WORDS 78

APPENDIX C TWO-LEVEL RULES **80**

1 Introduction

This thesis describes the morphology of Esperanto by a two-level morphology system. Esperanto is an agglutinating language, therefore the two-level morphology approach is extremely suitable for it.

In Esperanto, there are no phonological alternations and nearly no irregularities, therefore, there is a tendency to think that the morphological analysis must be very easy. The problem is that the word building is very rich, consisting of many short morphemes and nearly everything is allowed. Therefore, some words can be analyzed by many different ways. This is mostly no problem for a human – with the knowledge of the world and of the context. I have tried to allow as much flexibility to the word building as possible, with some restrictions of surely impossible combinations. To achieve this goal, I have used a mixture of linking lexicons and using two-level rules. I have implemented most of the areas of the Esperanto word building, including words created by affixes, classical composites, abbreviations, and much more. On contrary, inflection is totally unambiguous, very simple and regular, therefore it was no problem to implement it.

Many words that could not be regarded as derived in other languages (at least from synchronic point of view) and would require separate lexical entries (*town – mayor – city*), are derived in Esperanto (*urbo – urbestro – urbego*). Therefore, the size of the lexicon can be substantially smaller.

Generally, two-level system can be used for both, generation and analysis. However, I have dealt only with analysis. The resulting system has lexicon with about 11 thousands entries. It was tested on set of Esperanto texts containing about 460 000 words and has covered about 97.5 % of them. Most of the unanalyzed words are proper names or misspellings. It was also shown that Esperanto is a language with very high lexical homonymy (13.6 %) – this is the price for its rich word building. There are still some areas to cover – especially proper names and their capitalization and connection to Esperanto inflection. Other question is adapting the system for using as a reasonable generator.

1.1 About Esperanto

Esperanto is the most commonly used artificial language. It was created by Polish physician Ludwig L. Zamenhoff and was first presented in 1887. The name of the language comes from the pseudonym (“Doktoro Esperanto”) used by the author in his first textbook.

Esperanto can be learned considerably quicker than a typical natural language. The grammar is extremely regular, yet not primitive. There is only one paradigm for nouns and one paradigm for verbs. There is a simple relation between written and spoken text. The word order is “free”, allowing topic-focus articulation.

About 70% of Esperanto vocabulary come from Romance languages, about 20% from Germanic languages and English and some part from Slavic languages. The word-building is very rich and highly regular.

Most of estimated numbers of Esperanto speakers range from 1 to 10 millions[1]. There is about 1 000 of native speakers. Several tens of thousands of books have been published in Esperanto (original and translated), and there are many periodicals.

There are two tendencies in the current Esperanto (as in any other language) – conservative and progressive. The conservative group uses as a measure of the correctness of the language the books written by Zamenhof, mostly from so called Fundamento[2]. They say that these things are untouchable; even Zamenhof’s mistakes. There is a second group trying to change the language to make it more international, more close to English, easier to use etc. Some of the proposals are unsuccessful, some are partially used and some are even made official by Akademio – the headquarter of the Esperanto world. There is a third group of Esperanto users – AIL, group of scientists that uses the language for pragmatic reasons and wants to distinguish itself from the first two groups. They call the language ILO – Internacia Lingvo (International language), the original name used by Zamenhof.

---

[1] Funk and Wagnall’s The World Almanac states two millions of speakers. (The World Almanac is a part of Microsoft Bookshelf 1994)

[2] Zamenhof, L. L: *Fundamento de Esperanto* – 9-a eldono, EFE, Marmande, France, 1963. This bibliographical information was taken from Wennergren: Plena Manlibro de Esperanta Gramatiko, 1989, (PMEG) – literatu.htm.

2 Esperanto grammar

Esperanto is a highly regular language of agglutinal type. Some of the categories are expressed synthetically and some analytically. There is only one paradigm for nouns and one paradigm for verbs.

2.1 Writing and pronunciation

Esperanto uses Latin alphabet with 28 letters:

A B C Ĉ D E F G Ĝ H Ĥ I J Ĵ K L M N O P R S Ŝ T U Ŭ V Z
a b c ĉ d e f g ĝ h ĥ i j ĵ k l m n o p r s ŝ t u ŭ v z

The pronunciation of letters without diacritics is nearly the same as the same letters in IPA (except *c*). Letter *c* is pronounced as *ts* in *hats*, *ĉ* as *ch* in *church*, *ĝ* as *g* in *geography*, *ĵ* as *s* in *vision*, *ŝ* as *sh* in *ship* and *ŭ* is used in diphthongs (*aŭ* – *ow* in *how*).

Six letters – *ĉ, ĝ, ĥ, ĵ, ŝ* and *ŭ* – are unique to Esperanto. In Esperanto, the diacritical mark over the first five letters is called *cirkumflekso* (circumflex); the diacritical mark over u is called *hoketo* (hacek). There are two main alternatives to these diacritical marks:

1) To use letter *h* instead of circumflex and drop hacek: *ch*, *gh*, *hh*, *jh*, *sh*, *u*.

This is the official alternative, which was proposed by the creator of the language. The advantage of it is that the transcribed word look more internationally: *shi* – *she*, *shipo* – *ship*, *chambro* – *room* (in French *chambre*), *automobilo* – *car*, *Europo* – *Europe*. The disadvantage consist in problematical converting from this transcription back to the alphabet with diacritical signs – there are few roots like *ekshibici*, *ghett*, etc. where *sh* and *gh* does not stand for *ŝ* and *ĝ*. In addition, of course, there is no difference between *u* and *ŭ*.[3]

2) To use letter *x* instead of circumflex and hacek: *cx, gx, hx, jx, sx, ux*.

This alternative is not official, but is widely used on WWW and other texts on computers. The advantage is that there is a direct mapping between words in it and in the alphabet with diacritical marks – letter *x* is not used in Esperanto. The disadvantage is that words are not so similar to western languages.

In my morphological analyzer, I am using the transcription with *x* – as was said it is easy to convert it to both other transcriptions.

The pronuntiation of *ĥ* is hard for people of some nationalities. The letter *ĥ* is also very rare. There is a tendency to replace this letter (*ĥemio* = *kemio* – *chemistry*, *teĥniko* = *tekniko* – *technique*). There is even a rule that any sequence *rĥ* can be replaced by *rk:* (*arĥitekto* = *arkitekto* – *architect*, *arĥeologo* = *arkeologo* – *archeologist*).

2.2 Case and number

Esperanto has no grammatical gender.

There are two numbers – singular and plural. Singular has zero ending, plural is expressed by the ending *j* added to the basic form of the word.

singular	*amiko*	*friend*
plural	*amiko**j***	*friends*

Regarding of the form, there are two cases – nominative and accusative. Nominative has zero ending, accusative is expresses by the ending *n* added to the word in proper number.

nom. sg.:	*bela domo*	*nice house*
acc. sg	*bela**n** domo**n***	*nice house*
nom. pl	*belaj domoj*	*nice houses*
acc. pl.	*belaj**n** domoj**n***	*nice houses*

2.3 Article

Esperanto has a definite article *la*. The article is alike for all cases and number. There is no indefinite article. The usage of the definite article is similar to the usage of it in western languages.

[3] In my opinion, this ambiguity is highly insignificant. There are only few roots where the pair *ch, gh,* etc. does not stand for ligature. More ambiguities can be created by word building, when the second morpheme starts with letter *h* (there is no such affix). However, ambiguity is so common in Esperanto, that this would increase the overall ambiguity of a word only a bit.

Final *a* of the article can be dropped and replaced by an apostrophe, if the article is preceded by a preposition ending with a vowel: *la amo de l'patrino*[H] *– the love of the mother*.

Rarely, mostly in poetry, the elision is done also in other cases if it is possible to pronounce it: *L' espero, l' obstino kaj la pacienco...*[4] *– Hope, stubbornness and patience...*

2.4 Noun

Substantives are formed by adding the ending *o* to the stem.

E.g.: *dom**o** – house, amik**o** – friend, arb**o** – tree, bird**o** – bird, te**o** – tea, tag**o** – day, radi**o** – radio, Eŭrop**o** – Europe, tabl**o** – table, bier**o** – beer, paper**o** – paper, bildkart**o** – post card*

This final *o* may be dropped and replaced by an apostrophe. The stress is not affected by it. The *o* can not be elided if it is followed by plural or accusative ending.

E.g.: *mia amik' – my friend*

2.4.1 Proper names

Proper names can be classified into three groups – totally assimilated, partly assimilated and original.

The form of totally assimilated names was transcribed to Esperanto alphabet, has Esperanto pronunciation and follows Esperanto grammatical rules (they have *o* ending).

Jakobo – James, Paŭlo – Paul, Ĝenevo – Geneva, Eŭropo – Europe, Javo – Java, Nov-Zelando – New Zealand, Maro Ruĝa – Red Sea, Prago – Praha, Prague, Ŝekspiro – Shakespeare

Female proper names are formed the same way (*Lukrecio – Lucretia, Mario – Maria*) or are created from male names by prefix *in* (*Paŭlino – Pauline, Juliino – Julia, Mariino – Maria*). Today the tendency is to keep original form as much as possible. The names are often without an ending (*Elizabet – Elisabeth*) or with the ending *a*, which is normally used for adjectives, is used: *Eva, Johana – Joan, Marta*.

The partially assimilated names use Esperanto alphabet (with or without Esperanto pronunciation) but do not have noun ending and the non-assimilated names keep their original form:

Beijing, Elizabeth, Eva, Allah, Nelahozeves, Praha, Goethe, Fuĵijama, Gorbaĉov, Shakespeare.

2.4.1.1 Declination of proper names

With assimilated names is no problem – they are declined as any other Esperanto word. If the unassimilated name ends with a vowel (pronounced), the accusative ending *n* is simply added to it (often preceded with hyphen to facilitate understanding): *Dante-n, Evan, Anna-n, Brno-n, Bordeaux-n*. If the name ends with a consonant (pronounced), the noun ending is added: *Bill-on Clinton-on, Lebanonon.*

Proper names are normally only in singular. However, there are few exceptions: *Andoj – Andes mountains, Filipinoj – Filipines, la Burbonoj – Bourbons.*

2.4.1.2 Capitalization

The basic rules for capitalization are the same as in the most of other languages – with capitals are written the proper names of persons, towns, rivers, countries, continents, books, etc. Very often also names of months, nations. For the rest of the rules, see PAG §39.

Petro – Peter, Napoleono, Clinton, Eŭropo – Europe, Kanado – Canada, Pasko – Easter, Dio – God, Allah, la Biblio – Bible, Plena Ilustrinta Vortaro – The Full Illustrated Dictionary, Junio – June, angloj or *Angloj – Englishmen,* etc.

2.5 Adjectives

2.5.1 Form

Adjectives are formed by adding the ending *a* to the stem.

E.g.: *amik**a** – friendly* (adj.), *blank**a** – white, kar**a** – dear, bon**a** – good, Eŭrop**a** – European*

[4] Cited from PMEG, however originally from Zamenhof, L: Fundamenta Krestomatio de la Lingvo Esperanto, p.300

2.5.2 Comparison

Comparison of adjectives is done analytically. Comparative is formed by *pli* + positive, superlative by *la plej* + positive.

positive	*bela*	*beautiful*
comparative	*pli bela*	*more beautiful*
superlative	*plej bela*	*the most beautiful*

Vi estas tiel bela kiel ŝi. – *You are so beautiful as she is.*
Vi estas pli bela ol ŝi. – *You are more beautiful than she is.*
Vi estas la pli bela el ni. – *You are the most beautiful from us.*

It is also possible to compare in opposite direction using prefix *mal* – then comparative is formed by *malpli* + positive, superlative by *malplej* + positive. It is equivalent to put prefix to the *pli*/*plej* and to the compared adjective.

positive	*bela*	*beautiful*
comparative	*malpli bela* = *pli malbela*	*less beautiful*
superlative	*malplej bela* = *plej malbela*	*the least beautiful*

Vi estas malpli bela ol ŝi. = *Vi estas pli malbela ol ŝi.* – *You are uglier than she is.*
Vi estas malplej bela ol ni. = *Vi estas plej malbela ol ni.* – *You are the ugliest from us.*

Adverbs compare the same way (see 2.10.2)

2.6 Pronouns

2.6.1 Personal pronouns

Singular:	nominative	accusative
1st person	*mi* – *I*	*min* – *me*
2nd person	*vi* – *you*	*vin* – *you*
3rd person	*li* – *he* (for male beings)	*lin* – *him*
	ŝi – *she* (for female beings)	*ŝin* – *her*
	ĝi – *it*	*ĝin* – *it*
Plural		
1st person	*ni* – *we*	*nin* – *us*
2nd person	*vi* – *you*	*vin* – *you*
3rd person	*ili* – *they*	*ilin* – *them*

N.B. Sex of 3rd person is expressed by different words, not by suffixoid *in* (*amik**in**o* – *she-friend*) or prefixoid *vir* (***vir**kato* – *tom cat*) as by nouns.

Like English, Esperanto uses the same pronoun in 2nd person for both numbers. If it is necessary to distinguish vouvoyer and tutoyer (*cidiri*), it is possible to use for 2nd person singular *ci* (thou, fra: tu, deu: du).

There is also reflexive pronoun *si*. *Si* is used instead of classical 3rd person pronouns (*li*, *ŝi*, *ĝi*, *ili*) when referring to the subject of the sentence:

Paŭlo lavas sin. – *Paul washes himself.*
Paŭlo lavas lin. – *Paul washes him.* (It means someone else).
Ili lavas sin. – *They wash themselves.*

For 1st and 2nd person are used classical pronouns (*Mi lavas min.* – *I wash myself*). *Si* can form accusative but nominative can be used only as prepositive nominative.

The prefix *mem* (*self*, *own*) is very often replaced by accusative of the pronoun *si*. However, it is better to look at it as a separate prefix. First, it should be nominative – the accusative form is used only because of the easier pronunciation. Second, it is used also for first and second person, in which case the form *mi*, *ni* or *vi* should be used.

General subject is expressed by pronoun *oni*. *Oni* is vague and can stand for one or more persons – so predicative can be in singular (*Oni devas ĉiam esti preta.*[M] – *It is necessary to be ready all the time.*) or in plural (*Oni estas maljustaj koncerne ilin.*[M] – *People are not fair to them.*). Singular is preferred. *Onin* is possible, but not used.

Pronoun *ĝi* should be used also when referring to a human without specifying its sex. Some Esperanto speakers have proposed a new pronoun *ri*, reserving *ĝi* for things and non-human beings. However, *ri* is used very rarely. Other forms (*liŝi*, *ŝili* and *ŝli*) were proposed too, but they are even more rare. Some other Esperanto speakers would like to have pronoun for female 3rd person plural – plural of *ŝi* – they have proposed *iŝi* (as analogy to the pair *li* – *ili*). This pronoun is also used only rarely.

2.6.2 Possessive pronouns

Possessive pronouns are formed from personal pronouns by adding the adjectival ending *a*. Possessive pronouns are declined as adjectives. The pronoun has to agree in number and case with the thing(s) that are possessed.[5]

mia domo – my house
miaj domoj – my houses
Mi vidas mian domon. – I see my house.
Mi vidas miajn domojn. – I see my houses.

Li vidas sian domon. – He sees his own house. (The house belongs to the person who sees it.)

Li vidas lian domon. – He sees his house. (The house belongs to some other person.)

It is possible also to form possessive form *oni* – *onia*, but I cannot find any interpretation for it.

2.7 Correlatives

Correlatives (*Korelativoj* or *Tabelvortoj*) is system of 45 words, partly pronouns (*kiu – who, tiu – this, kies – whose,* etc), partly adverbs (*kie – where, tie – there, kiom – how much,* etc).[6]

2.7.1.1 System of correlatives

Every correlative consists of two parts – first and second.
First parts[7]:

ki- = interrogative (*demandovorto*)
ti- = demonstrative (*montrovorto*)
i- = indefinite (*nedifinita vorto*)
ĉi- = universal (*tutampleksa vorto*)
neni- = negation (*nea vorto*)

Second parts:[8]

-*u* = individuality (*individuo*)
-*o*[9] = thing (*aĵo*)

---

[5] See chapter 2.14.2 Agreement 1)

[6] In other languages is very often also some system of some adverbs or pronouns, but mostly it is not so regular and complex as in Esperanto. See for example English (Source: J.M.D. Meiklejohn, The English Language - Its grammar, history and literature, 1895):

Pronoun	Place In	Place To	Place From	Time In	Manner	Cause
Wh-o	Whe-re	Whi-ther	Whe-nce	Whe-n	Ho-w	Wh-y
Th-e	The-re	Thi-ther	The-nce	The-n	Th-us	Th-e
He	He-re	Hi-ther	He-nce			

[7] There are also unofficial forms with the first part *al – another (aliu – somebody else, aliel – in another way, etc.)* This set was created by analogy from the word *alia – another*. The words derived from the official root and the unofficial set of correlatives have different meaning.

[8] Some of the second parts are same as normal endings, but they have different meaning – ordinary *u* stands for volitive, ordinary *e* stands for any adverb, not only for place, ordinary *a* stands for any adjective, not only for quality. Only *o* has nearly the same meaning.

[9] This *o* has nothing to do with the noun ending *o*, so it is impossible to replace it by apostrophe.

-*a* = quality (*kvalito*, *eco*)
-*es* = possessor (*posedo*)
-*e* = place (*loko*)
-*am* = time (*tempo*)
-*al* = cause (*kaŭzo*)
-*el* = manner (*maniero*)
-*om* = quantity (*kvanto*)

By combining of these two sets, it is possible to form 45 words:

	interrogative *demanda*	demonstrative *montra*	indefinite *nedifina*	universal *kolektiva*	negative *negativa*
individual. *individuo*	*kiu* *who* *which*	*tiu* *that one* *that*	*iu* *somebody,* *some*	*ĉiu* *everybody* *every, all*	*neniu* *nobody* *no, none*
thing *neŭtraĵo*	*kio* *what*	*tio* *that thing*	*io* *something*	*ĉio* *everything*	*nenio* *nothing*
quality *kvalito*	*kia* *what kind of*	*tia* *that kind of*	*ia* *some kind of*	*ĉia* *every kind of*	*nenia* *no kind of*
possession *posedo*	*kies* *whose*	*ties* *that one's*	*ies* *someone's*	*ĉies* *everyone's*	*nenies* *nobody's*
place *loko*	*kie* *where*	*tie* *there*	*ie* *somewhere*	*ĉie* *everywhere*	*nenie* *nowhere*
time *tempo*	*kiam* *when*	*tiam* *then*	*iam* *sometime*	*ĉiam* *always*	*neniam* *never*
cause *kaŭzo*	*kial* *why*	*tial* *so*	*ial* *for some reas.*	*ĉial* *for every reas.*	*nenial* *for no reas.*
manner *maniero*	*kiel* *how*	*tiel* *thus*	*iel* *somehow*	*ĉiel* *in every way*	*neniel* *in no way*
quantity *kvanto*	*kiom* *how much*	*tiom* *so much*	*iom* *some*	*ĉiom* *all of it*	*neniom* *no amount*

2.7.1.2 Declination of correlatives

Correlatives of individuality (-*iu*) can form accusative and plural.
Correlatives of things (-*io*) can form accusative, but normally do not form plural.

*Kio**n** tiu**j** homo**j** ĉio**n** ne elpensas. – What all do the people think out!*

Adverbial correlatives of place (-*ie*) can form accusative to mark direction.

Mi estas tie. – I am there.
*Mi iras tie**n**. – I am going there.*

Other adverbial correlatives (time – -*am*, cause – -*al*, manner – -*el*, quantity – -*om*) and possessive correlatives (-*ies*) do not decline.

2.7.1.3 Using parts of correlatives in word building

It is not normal to use first or second part of correlative and combine it alone with some root or affix. However, there are some few exceptions – *neni-aĵo*, *neni-eco*, *neni-igi*, *neni-iĝi*, *ti-aĵo*.[10]

Neniaĵo – nearly nothing, thing with no value
Viaj kontraŭuloj fariĝos neniaĵo kaj pereos.[11] *– Your enemies will do nearly nothing to themselves and they will perish.*
Nenieco – quality to be like nothing, nothingness
dezerta regno de la nenieco[12] *– desert kingdom of nothingness*
Neniigi – destroy
Mi neniigos vin, kaj vi ne plu ekzistos[13] *– I will destroy you and you will no more exist.*
neniiĝi – disappear
Li disneniiĝis kiel fumo. – He disappeared like a puff of smoke.
tiaĵo – thing of that quality (*ti-aĵo*, from *tia* and *aĵo*)

[10] PMEG – taqord.htm – down
[11] From PMEG, originating from Old Testament translated by L. Zamenhof
[12] From PMEG, originating from Schiller, F.: *La Rabistoj*, translated by L. Zamenhof
[13] From PMEG, originating from Old Testament translated by L. Zamenhof

Mi ne ŝatas tiaĵojn. – *I don't like things that look, behave, etc like that.*

2.7.1.4 Using correlatives in word building

Some correlatives can accept different category endings, some can accept suffixes and some can even form composites with other roots. Very often is the set of possible derived forms restricted to some few traditional forms. I will go through one type of correlatives after another.

Individual – -iu

The individual form can be before nearly any root (see *–ia*):

tiumomente – in that moment, tiunokte – during that night, tiusence – in that sense, kiusence – in what sense, iusence – in some sense, etc.

It is impossible to add any ending or true suffix to it. Of course, it can be declined.

Quality – -ia

The correlatives of quality can be also before nearly any root. The difference usage of *-iu* and *–ia*, is implied by their meaning – *-iu* refers to some concrete thing, occasion, etc., *-ia* refers to some type, quality or style of thing, occasion, manner etc. Sometimes it is hard to distinguish these two groups.

tiamaniere – in such manner, tiasence – in such sense, tiaspeca – of such type, similar, analogous, kiamaniere – in what manner, how, ĉiamaniere – in all manner

Place – -ie

It is possible to form adjectives from correlatives of place (*-e*), e.g. *tiea*, *ĉiea*, etc., with meaning "finding itself there, everywhere, etc."[14]. These adjectives are normally declined.

La ĉiea pluvo detruis ĉiuj vojojn. – Rain that was everywhere destroyed all roads.

Sometimes, it is also possible to see form *tieulo – the man from there*.

Quantity – -iom

Correlatives or quantity can have ordinal form by adding the ending *a*, e.g.: *kioma*, *tioma*.

Sur la kioma etaĝo vi loĝas? – On which floor do you live?

It is also possible to diminish or augment the quantity (practically only *iom*): *iomete – a bit of, iomege – some large quantity, iometo – a bit.* Some numerical suffixes can also be added: *kiomoble – how many times, kiomfoje – how often*

The forms with adverbial *e* are only emphasized forms of the original: *iome*.

Time – -iam

The forms with adjectival ending (*-a*), e.g. *tiama*, *ĉiama. iama*, etc., with meaning "existing in that time, existing always, existing in some time (in the past).

Cause – -ial

The only derived word is *kialo – the reason, motive*.

Manner – -iel

Tiele and *iele* are emphasized forms of *tiel* and *iel*.

The forms with adjectival ending (*-a*), e.g. *tiela*, *kiela* etc. are equivalents to *tiamaniera*, *kiamaniera*, etc.

It is also possible to see word **tielmaniere* – the correct form is *tiamaniere* or *tiumaniere*.

Possession – -ies, Thing – -io

I do not know about any derived forms.

2.8 Numerals

2.8.1 Cardinal numerals

There are 23 elementary cardinal numerals (*bazaj numeraloj*):

nul – 0
unu[15] – 1

[14] PMEG – el_e.htm, subchapter Vortfarado

du – 2
tri – 3
kvar – 4
kvin – 5
ses – 6
sep – 7
ok – 8
naŭ – 9
dek – 10
cent – 100
ok – 8
naŭ – 9
dek – 10
cent – 100
mil – 1000

Other cardinal numerals are formed by combining of these elementary ones:

dek unu – 11
dek tri – 13
dudek – 20
dudek kvin – 25

tridek – 30
okdek – 80

cent kvin – 105
cent tridek ok – 138
naŭcent – 900

mil ducent kvardek sep – 1 247
tri mil – 3 000
dudek mil – 20 000
cent okdek unu mil kvarcent naŭdek tri – 181 493

Numerals 1 to 999 999 can be expressed by the following formula N:

N = [[I'] * *mil*] + [I]
I = [[*du*-*naŭ*] * *cent*] + [[*du*-*naŭ*] * *dek*] + [*unu*-*naŭ*]
I' = [[*du*-*naŭ*] * *cent*] + [[*du*-*naŭ*] * *dek*] + [*du*-*naŭ*]

- [x] means that x is optional, with one exception – the resulting string of the whole formula N cannot be empty.
- *du*-*naŭ* or *unu*-*naŭ* means one of the numerals between *du* and *naŭ*, or *unu* and *naŭ*.
- Parts separated by * are written together in I and I', and are written with space in between in N. Parts separated by + are written with space in between.

Interpretation of the resulting numeral is following: Elementary numerals are replaced by numbers and + and * are treated as classical arithmetic operators.

These cardinal numerals can be considered as nondeclinable adjectives. Counted things are normally declined (and in plural, if the numeral is different then *unu*).[16]

---

[15] In special occasions (counting things) it is possible to use *un'* instead of *unu*. (PMEG – asqnor.htm):

Un'! du! un'! du! - La soldatoj marŝis.[M] *– One, two, one, two – soldiers marched.*
"Un', du, tri, kvar", li kalkulis laŭte.[M] *– "One, two, three, four", he counted loudly.*

However, it is impossible to use that form in normal sentences: **Mi havas nur un' amikon.*[M] *– I have only one friend.*

[16] There is no rule about number of the counted thing after *nul*. PMEG (jfquzasp.htm) says that it is preferred to say *neniu*, if it is possible: *Post tio restos nul homo(j).* = *Post tio restos neniu(j) homo(j).* *– After that nobody will stay.*, *Mi aĉetis nul komo kvin kilogramo(j)n da rizo.* *= I have bought 0.5 kilograms of rice.*

*unu viro – one man, kvin viro**j** – five men*
***Kvin** amiko**j** iras en arbaron. – Five friends go to the forest.*
*La instruisto laŭdas **kvin** lernanto**jn**. – The teacher praises five pupils.*

The numeral *unu* can be in a function of a pronoun. In that case, it is declined.
Unuj legis, kaj aliaj skribis. – Ones have been reading and the others have been writing.
Estas facile unujn ami kaj aliajn malami.[H] *– It is easy to love ones and to hate others.*

There are different opinions about accusative form of pronoun *unu* in singular. Some authorities[17] say that it is not correct to add accusative ending *n* with *unu* in pronominal function. The reason, why they do not want to allow it, is that it is hard to distinguish between numeral and pronominal *unu* (in contrary to the form *unuj*, which can be only pronoun). Some authorities[18] are not so strict and just say that the absence of *n* in accusative of pronoun *unu* is illogical, and has no other than historical reasons. Some few normally use form *unun*.

Numerals as *miliono – 10^6*, *miliardo – 10^9*, *biliono = duiliono – 10^{12}*, etc. are nouns[19] and are normally declined:
*Mi havas unu milion**o**n. – I have one million.*
*Mi havas dek milion**o**jn. – I have ten millions.*
and counted objects are in prepositive using preposition *da*:
Cent milionoj da dolaroj. – Hundred millions of dollars.
There is no strict rule about mixed expression (noun numerals with pure numerals)[20]:
*Li havas dek milion**o**jn tricent mil naŭcent sepdek ok **da** dolaroj. = Li havas dek milion**o**jn tricent mil naŭcent sepdek ok dolaro**j**n. – He has $10 300 978.*

2.8.2 Non cardinal numerals[21]

Other than cardinal numerals are formed by suffixes and endings added to the last part of the cardinal numeral. The spaces between parts of the cardinal numeral are replaced by hyphen.
trimil okcent dudek kvin – 3 825
trimil-okcent-dudek-kvina – 3 825th

2.8.2.1 Ordinal numerals

Ordinal numerals (*Ordaj numeraloj*) are formed by adding the adjective ending *a*:
*unu**a** – first*
*du**a** – second*
*mil-kvincent-sesdek-tri**a** – 1563rd*
You can also form these ordinal numerals from numeral nouns by replacing the noun ending by the adjective ending: *milion**o*** → *milion**a***
*Nia milion**a** kliento ricevos specialan donacon.*[M] *– Our millionth client will receive special present.*

Ordinal numerals are normally declined as adjectives:
Mi skribas trian ĉapitron. – I am writing the third chapter.

2.8.2.2 Adverbial numerals

Adverbial numerals (*Numeralaj adverboj*) are formed by adding the adverbial ending *e*:
*unu**e** – for first time, first (in a list)*
*du**e** – for second time, second (in a list)*
*mil-kvincent-sesdek-tri**e** – for 1563rd*

---

[17] See (PMEG) – nvqbazun.htm
[18] See Kalocsay, Waringhien: Plena Analiza Gramatiko de Esperanto (PAG), 1985, §64
[19] In the past, also *nul* was only a noun – *nulo*, today a form *nul* is normally used, and form *nulo* is used for the name of the number – the same relation as between *kvin – kvino*, *dek – deko*.
[20] See also PMEG – nvqmik.htm
[21] See also PAG §87A-F, PMEG nvq.htm – Nomboroj

2.8.2.3 Names of numbers

Names of numbers are formed by adding the noun ending *o*:

*unu**o*** – *number one*
*du**o*** – *number two*
*cent**o*** – *number hundred*
*kvincent-tridek-sep**o*** – *number five hundred thirty seven*

2.8.2.4 Multiplication numerals

Multiplication numerals (*multiplikaj numeraloj*) are formed by suffix *obl*.

*tri**obl**a* – *three as much in size, strength, number, or amount*
*tri**obl**e* – *three times*
*tri**obl**o* – *a number or quantity three times as great as another*
*tri**obl**igi* – *to make something three times bigger, larger, etc.*

or by suffix *foj*:

*tri**foj**a* – *occurring three times*
*tri**foj**e* – *three times*
*tri**foj**o* – *an occurrence three times*
*tri**foj**igi* – *to make something occurring three times*

There is a difference between *obl* and *foj*: The former means multiplication, the latter repetition.

duobla pago – *salary two times as big as normal*
dufoja pago – *salary paid two times to the same person*

2.8.2.5 Collectives

Collectives (*kolektivigaj numeraloj*) are formed by suffix ***op***:

*du**op**a* – *having groups of two*
*du**op**e* – *in groups of two*
*du**op**o* – *group of two, a pair*
*kvar**op**o* – *quartet*
*marŝi kvar**op**e* – *march in groups of four*

2.8.2.6 Fractions

Fractions (*frakcioj*) are formed by suffix ***on***:

*du**on**o* – *half*
*du**on**a* – *being a half*
*du**on**e* – *to the extend of one half*
*du**on**igi* – *to halve*

Counted objects after fractions are connected with preposition *de* and not *da*.

Duono de ni mortos.[A] – *One half of us died.*
triona horo = triono de horo – *three quarters of an hour*
Li faris sian taskon nur trione.[A] – *He did only one third of his task.*
La tanko estas duone malplena. – *The tank is half-empty.*

Nominal fractions are normally declined:

Li donis al mi duonon de sia pano.[M] – *He gave me half of his bread.*
Li trinkis duonan litron da lakto.[M] = *Li trinkis duonon de litro da lakto.*[M] – *He drunk one half of the liter of the milk.*

2.8.2.7 Distribution

Distribution[22] (*distribuo*) of objects is expressed by preposition *po*:

La gastoj trinkis vinon po du glasoj.[A] – *Each guest drunk two glasses of wine.*
La gastoj venis po tri. – *Guests came in groups of three.*
Ili ricevis po kvin pomojn. – *They received five apples each.*
Prenu la medikamenton po 20 gutoj. – *Each time you use the medicine, take 20 drops.*

From *po*, you can form also adjective – *poa*, and adverb – *poe*:

Ili ricevis poan korbon da pomoj.[A] – *They received a basket of apples each.*

[22] See also PMEG rv_po.htm, PAG §87C

Tiuj studentinoj havis poe plurajn amantojn.[A] *– These girl-students had more lovers each.*

These forms are rare, because *poa* = *po unu*, and *poe* = *po*.

2.8.2.8 Other topics

I have not covered many topics that are not important for morphology – how to express dates and times, problems of using numerals with some prepositions, mathematical expressions, etc. These topics are thoroughly covered in PAG §87G & I or PMEG – nvq.htm

2.9 Verb

Esperanto has most of the verb forms found in western languages, and some more. All forms are regular. The forms used more often are created synthetically, the rest is created analytically using auxiliary verb *esti* – to be and participles.

I will show first the simple forms and then complex forms. See also suffixes *ig* and *iĝ* in chapters 3.2.2.1 and 3.2.2.2. Because of the purpose of this grammar overview, I will not spend time with describing all aspects of usage of all these various forms. This topic is covered in PAG §110-119.

2.9.1 Infinitive

Infinitive (*infinitivo*) is formed from stem by adding the ending *i*.

*est**i** – to be, sid**i** – to sit, kapt**i** – to catch, martel**i** – to hammer, skrib**i** – to write, bezon**i** – to need.*

Infinitive has the same meaning as in many other languages:

Mi ĝojas vin vidi. = Mi ĝojas, ke mi vin vidas. – I am happy to see you.
Mi vidis la knabon kuri. = Mi vidis, ke la knabo kuras. – I see the boy running.
Kritiki estas facile, sed fari estas malfacile. – It is easy to criticize, but it is hard to work.
Mi povas legi. – I can read.

2.9.2 Vowels of tense

All tenses (relative and absolute) are formed uniformly using three vowels:

a for present tense, contemporaneous, imperfectness
- *mi kapt**a**s – I catch*
- *kapt**a**nta – catching*
- *kapt**a**ta – being caught*

i for past tense, anteriority, perfectness
- *mi kapt**i**s – I caught*
- *kapt**i**nta – having caught*
- *kapt**i**ta – having been caught*

o for future tense, succession, intention
- *mi kapt**o**s – I will catch*
- *kapt**o**nta – going to catching*
- *kapt**o**ta – going to be caught*

2.9.3 Indicative

Indicative (*indikativo*) is formed by adding a vowel expressing tense followed by *s* to the stem:

present	*mi kapt-**a**-s*	*I catch*
past	*mi kapt-**i**-s*	*I caught*
future	*mi kapt-**o**-s*	*I will catch*

The form of the verb is the same for all persons:

mi kaptas	*I catch*	*ni kaptas*	*we catch*
vi kaptas	*you catch*	*vi kaptas*	*you catch*
li kaptas	*he catches*	*ili kaptas*	*they catch*

2.9.4 Conditional

Conditional (*kondicionalo*) is formed by adding the ending *us* to the stem.

*mi kapt-**us** – I would catch*

2.9.5 Imperative

Imperative (*volitivo*) is formed by adding the ending *u* to the stem:

*kapt-**u** – catch*
*ni kapt-**u** – let's catch*

Volitive is used also as subjunctive:

Mi petas, ke li venu.[H] *– I ask that he comes.*
Mi deziras al vi, ke vi resaniĝu.[H] *– I wish you to get healthy again.*
Mi alportis la libron, por ke vi ĝin tralegu.[H] *– I brought the book for you to read.*

2.9.6 Participles, Gerunds, Verbal nouns

2.9.6.1 Participles

There are three types of active and three types of passive participles – present, past (perfect) and future (predicative).

Active participle is created from the stem by adding a vowel of tense, followed by *nt*, followed by adjective ending *a*.

present	*kapt-**a**-**nt**-**a***	*catching*
past	*kapt-**i**-**nt**-**a***	*having caught*
future	*kapt-**o**-**nt**-**a***	*going to catch*

Passive participle is created from stem by adding a vowel of tense, followed by *t*, followed by adjective ending *a*.

present	*kapt-**a**-**t**-**a***	*caught, being caught*
past	*kapt-**i**-**t**-**a***	*having been caught*
future	*kapt-**o**-**t**-**a***	*going to be caught*

2.9.6.2 Gerunds

Gerunds are formed from participles by replacing the adjective ending *a* by the adverbial ending *e*.

Active gerunds:

present	*kapt-**a**-**nt**-**e***	*catching*
past	*kapt-**i**-**nt**-**e***	*having caught*
future	*kapt-**o**-**nt**-**e***	*going to catch*

Promentante ili kantas.[H] *– Walking, they are singing.*
Reveninte hejmen, ŝi komencis legi.[H] *– Having came home, she started to read.*
Pagonte li foriris.[H] *– He left before paying.*

Passive gerunds:

present	*kapt-**a**-**t**-**e***	*caught, being caught*
past	*kapt-**i**-**t**-**e***	*having been caught*
future	*kapt-**o**-**t**-**e***	*going to be caught*

Persekutate ili saltis en riveron.[H] *– Being persecuted, they jumped into the river.*
Kaptite ŝi vane provas liberiĝi.[H] *– Having been caught, she is trying to free herself.*
Jam kaptote, li eskapis.[H] *– Nearly caught, he escaped.*

2.9.6.3 Verbal nouns

Verbal nouns are formed from participles by replacing the adjective ending *a* by the noun ending *o*.

Active verbal nouns:

present	*kapt-**a**-**nt**-**o***	*the one who is catching*
past	*kapt-**i**-**nt**-**o***	*the one who is having caught*
future	*kapt-**o**-**nt**-**o***	*the one who is going to catch*

vojaĝanto – one who travels, voyager, lernanto – one who learns, pupil, aŭskultanto – one who listens, listener, abonanto – one who subscribes something, subscriber, vizitanto

– one who visits, visitor; *mortinto* *– one who died, the deceased,* *savonto* *– one who will save, savior, messiah,* *parolanto* *– one who speaks, speaker*

Passive verbal nouns:

present	*kapt-**a**-t-**o***	*the one who is being caught*
past	*kapt-**i**-t-**o***	*the one who is having been caught*
future	*kapt-**o**-t-**o***	*the one who is going to be caught*

sendito – one who was sent, messenger, jugoto – one who will be judged

2.9.6.4 Verbalized participles

"Verbalized participles" are formed from participles by replacing the adjective ending *a* by the verbal ending *i* and using it as a normal verb (of course not forming participle). These forms are quite rare, they are equivalents of complex verbal forms.[23]

kaptanti = esti kaptanta – to be (in state of being) catching
mi kaptintus = mi estus kaptinta – I would have caught

2.9.7 Complex verbal forms

Complex verbal forms are created using the auxiliary verb *esti* + participle. This way are expressed secondary active tenses, passive voice, and nuances of conditional and imperative.

These complex forms are not so often used. Very often if you use complex form in English you can use simple form in Esperanto.

2.9.7.1 Imperfect

Imperfect (*Imperfekto*) is expressed by the auxiliary verb *esti* + active present participle.

present	*mi estas kaptanta*	*I am catching*
past	*mi estis kaptanta*	*I was catching*
future	*mi estos kaptanta*	*I will be catching*

Imperfect is used when you want to express that the process was occurring in the same time as another process or that the process was continuous.

Li mortis. – He died.
Li estis mortanta. – He was dying.

Imperfect is not so often as English progressive tense, because it is often possible to use nonmarked simple verbal form.

2.9.7.2 Perfect

Perfect (*perfekto*) is expressed by the auxiliary verb *esti* + active past participle.

present	*mi estas kaptinta*	*I have caught*
past	*mi estis kaptinta*	*I had caught*
future	*mi estos kaptinta*	*I will have caught*

Perfect is used when you want to express that the process was already finished before some point in the present, past, or future.

2.9.7.3 Predicative

Predicative (*predicativo*) is expressed by the auxiliary verb *esti* + active future participle.

present	*mi estas kaptonta*	*I am going to catch*
past	*mi estis kaptonta*	*I was going to catch*
future	*mi estos kaptonta*	*I will be going to catch*

Predicative is used when you want to express that the process was going to happen after some point in the present, past, or future. Predicative is very often replaced by modal verbs with infinitive:

Mi estas kaptonta. – I am going to catch.
Mi volas/devas/intencas kapti. – I want to/must/am going to catch.

[23] See chapter 2.9.7.

2.9.7.4 Infinitive complex forms

Complex active infinitives are formed from the infinitive of the auxiliary verb *esti* + active participle.

imperfect	*esti kaptanta*	*to be (in state of being) catching*
perfect	*esti kaptinta*	*to have caught*
predicative	*esti kaptonta*	*to be going to catch*

2.9.7.5 Conditional and Imperative active complex forms

More precise forms of conditional or imperative can be expressed by combining of the auxiliary verb *esti* in simple form conditional/imperative with active participles.

Conditional

imperfect	*mi estus kaptanta*	*I would be catching*
perfect	*mi estus kaptinta*	*I would have caught*
predicative	*mi estus kaptonta*	*I would be going to catch*

Imperative

imperfect	*estu kaptanta*	*be catching!, You be catching*
perfect	*estu kaptinta*	*You have been/were catching*
predicative	*estu kaptonta*	*You shall catch*

2.9.7.6 Passive voice

Passive voice (*pasiva voĉo*) is expressed by the auxiliary verb *esti* + passive participle.

Imperfect

present	*mi estas kaptata*	*I am (being) caught*
past	*mi estis kaptata*	*I was (being) caught*
future	*mi estos kaptata*	*I will be (in state of being) caught*

Perfect

present	*mi estas kaptita*	*I have been caught*
past	*mi estis kaptita*	*I had been caught*
future	*mi estos kaptita*	*I will have been caught*

Predicative

present	*mi estas kaptota*	*I am going to be caught*
past	*mi estis kaptota*	*I was going to be caught*
future	*mi estos kaptota*	*I will be going to be caught*

Infinitive

imperfect	*esti kaptata*	*to be (in state of being) caught*
perfect	*esti kaptita*	*to have been caught*
predicative	*esti kaptota*	*to be in state of going to be caught*

Conditional

imperfect	*mi estus kaptata*	*I would be caught*
perfect	*mi estus kaptita*	*I would have been caught*
predicative	*mi estus kaptota*	*I should be caught*

Imperative

imperfect	*estu kaptata*	*Be caught!, You be caught*
perfect	*estu kaptita*	*Be caught!, You have been/were caught*
predicative	*estu kaptota*	*You shall/should be caught,*

Passive voice is very often expressed by different, simpler means:

Topic-Focus articulation:

La kato estis persekutata de la hundo. = *La katon persekutis la hundo.* – *The cat was chased by the dog.*

General subject:

La cervo estis pafita. = *Oni pafis la cervon.* – *The deer was shot.*

2.10 Adverbs

2.10.1 Form

Derived adverbs are formed by adding the ending *e* to the stem.

E.g.: *bon**e** – well, kuŝe – lying, facile – easily, rapide – fast, quickly, ekzakte – exactly, blue – bluely, sabate – on Saturday, nokte – in the night, skribe – in writing*

Many of them are derived from prepositions:

antataŭ – before → antataŭe – ahead
apud – beside → apude – nearby (adv.)
dum – during → dume – in the meantime
ekster – outside of → ekstere – outside
kontraŭ – against → kontraŭe – vice versa, conversely
kun – with → kune – together
post – after → poste – afterward
sub – under → sube – down
super – above → supere – above (adv.).

Many of the adverbs are not derived, so called primitive.

EG: *nun – now, jam – already, ĉi – near to me, ankoraŭ – still, baldaŭ – soon, hodiaŭ – today, tuj – immediately, plu – more, tre – very, ankaŭ – also, do – thus, nur – only*

Some adverbs are part of the system of so called correlatives (see chapter 2.7)

2.10.2 Comparison

Comparison of adverbs is done analytically. Comparative is formed by *pli* + positive, superlative by *plej* + positive.

positive	*bone*	*well*
comparative	*pli bone*	*better*
superlative	*plej bone*	*best*

Mi faras tiel bone kiel vi. – I work as good as you.
Mi faras pli bone kiel vi. – I work better than you do.
Mi faras plej bone el ni. – I work best from us.

It is also possible to compare in opposite direction using prefix *mal* – then comparative is formed by *malpli* + positive, superlative by *malplej* + positive. It is equivalent to put prefix to the adverb *pli*/*plej* and to the compared adverb.

positive	*bone*	*well*
comparative	*malpli bone*	*less well*
superlative	*malplej bone*	*the least well*

Vi estas malpli bela ol ŝi. = Vi estas pli malbela ol ŝi. – You are uglier than she is.
Vi estas malplej bela ol ni. = Vi estas plej malbela ol ni. – You are the ugliest from us.

Adjectives compare the same way (see 2.5.2)

2.10.3 Inflection

Adverbs of place can form accusative to mark the direction.

hejme – at home (place): *Mi estas hejme. – I am home.*
*hejme**n** – at home* (direction): *Mi iras hejmen. – I go home.*
kie – where (place): *Kie vi estas? – Where are you?*
*kie**n** – where* (direction): *Kien vi kuras? – Where do you run?*

2.11 Prepositions

Each preposition in Esperanto has its own fixed meaning and only few of them are used for more relations (mostly in temporal and local meaning and preposition *de*).

E.g. *al – to, anstataŭ – instead of, antaŭ – before, apud – next to, ĉe – near, ĉirkaŭ – around, da – of* (with quantity), *de – of, dum – during, ekster – outside of, el – from within, en – in, ĝis – till, inter – between, among, kontraŭ – against, krom – besides, kun – with, malgraŭ – in spite of, per – per, po – at rate of, por – for, post – after, pri – about, pro – for, because of, sen – without, sub – under, super – above, sur – on* (position), *tra – through, trans – across*

E.g. *al Praha – to Prague, en la ĝardeno – in the garden, per la martelo – using the hammer, sur la tablo – on the table, de la patro – from father*

In cases where no existing prepositions can be logically used, preposition *je* should be used. Preposition *je* has no concrete meaning.

E.g. *je la tri horo*[H] *– at three o'clock, krei je Dio – trust in God, esti je kvar jaroj pli aĝa – to be four years older*

Except these simple prepositions, there are also complex prepositions (*prepoziciaĵoj*). They are formed mostly by adverb and another preposition.

E.g. *dank' al*[24] *– by virtue of, thanks to, proksime al – near to, rilate al – relating to, kompare kun – comparing with, kontraste kun – in contrast with, kune kun – together with,* etc.

2.11.1 Types of prepositive

1) nominative-prepositive – preposition + noun in nominative

This is the most often case, e.g. *sur la tablo – on the table*

2) prepositive-prepositive – preposition + preposition + noun

This case is used, when it is necessary to express some move (spatial or temporal) from, to, over, etc. some place already expressed by some preposition. It is necessary to use two prepositions. The second preposition makes a prepositive with the noun and the first relates to this whole phrase. The second preposition expresses a position (spatial or temporal) and the first express an move relative to the complex of the second preposition together with the noun. E.g. *el sub la lito – from under the bed*

3) accusative-prepositive – preposition + noun in accusative

The accusative in this case is used to express a move[25]. It can be replaced by second case using preposition *al*: ***sub** la lito**n** = **al sub** la lito – under the bed* (direction).

*Li iras **en** la domo**n**. – He goes to the house.*

*Veturi **ekster** la urbo**n**. – Drive out of the town.*

It is beyond the scope of this grammar overview to explain meaning and usage of all prepositions of Esperanto, as a good source, I can recommend PMEG – maqvor.html and PIV.

2.12 Conjunctions

In Esperanto, as in other languages, there are coordinating (*konjunkcioj*) and subordinating (*subjunkcioj*) conjunctions.

Examples of coordinating are:

kaj – and, aŭ – or, sed – but, eĉ – even, nek … nek – neither … nor, etc.

Examples of subordinating are:

ke – that, se – if, ĉar – because, kvankam – although, etc.

It is also possible to use relative pronouns.[26]

There is nothing interesting on conjunctions from morphological point of view. For more information see PAG §123 and §124 or PMEG – kvq.html

2.13 Interjections

Interjections are not interesting from morphological point of view – they are just list of various sounds, screams, etc. like: *aha, huŝ, help, stop, brr, hura, uff, puf, miaŭ,* etc.

2.14 Short syntax overview

2.14.1 Accusative

It is possible to create accusative from nouns, adjectives, pronouns and adverbs of place. Accusative is formed by the ending *n* and has following functions:

1) Direct object.

Accusative marks direct object in the sentence.

*La knabo**n** mordis la hundo. – The boy was bit by the dog.*

[24] This is the only case of elision of an adverbial *e*.

[25] For overview of accusative usage, see chapter 2.14.1.

[26] See chapter 2.7 Correlatives.

2) Motion toward

Accusative marks motion toward, in contrary to position. It is not used after preposition *al* – to and *ĝis* – up to, because they can show only direction. It is possible to omit preposition *en* and to use pure accusative (but recommended only with names of cities and countries).

*Mi veturas (en) Londono**n**.* = *Mi veturas al Londono.* – *I go to London.* vs. *Mi estas en London.* – *I am in London.*

*Mi promenis en la ĝardeno**n**.* – *I walked into the garden.* vs. *Mi promenis en la ĝardeno.* – *I walked in the garden.* (I was already in the garden, walking around)

*Mi iras tie**n**.* – *I go there.* vs. *Mi estas tie.* – *I am there.*

*Mi iras hejme**n**.* – *I go home.* vs. *Mi estas hejme.* – *I am home.*

*La vagonaro kuras de Hanovero Berlino**n**.*[(FK.209)] – *The train goes from Hanover to Berlin.*

3) Date

*Hodiaŭ estas la dua**n** de julio.* – *Today, it is July 2.*

4) Time interval

*Li laboras tuta**n** tago**n**.* – *He works whole day.*

5) Weight, price, measure

*Mi pezas okdek kilogramo**j**n.* – *I weight 80 kg.*

*Ĝi kostas dek dolaro**j**n.* – *It costs $10.*

*La vojo estas longa cent kilometro**j**n.* – *The way is 100 km long.*

2.14.2 Agreement

As it was stated before, Esperanto has no grammatical gender, but has number and case. We can see two types of agreement.

1) Agreement inside of a nominal phrase. – All participants of a nominal phrase have to agree in case and number.

*Mia**j** bela**j** hundo**j** kuras en la ĝardeno*. – *My nice dogs are running in the garden.*

*Mi havas granda**n** domo**n**.* – *I have a big house.*

*Mi havas granda**jn** domo**jn**.* – *I have big houses.*

2) Agreement with subject. – Simple form of the verb is the same for all persons, but predicative adjective has to agree with subject (including participle in complex verb tenses):

*La studentoj estas diligenta**j**.* – *The students are diligent.*

*Ni estas kaptanta**j**.* – *We are catching.*

2.14.3 Word order

Word order in Esperanto is so called “free word order”. The word order in Esperanto is used to distinguish topic and focus – to express dynamics of the sentence[27] and not to distinguish between syntactical units (subject and object are distinguished by nominative and accusative). In English, it is necessary to use different means to express the same thing (passive, particles, relative sentence, etc.)[28]:

Kiun mordis la hundo? – *Who was bit by the dog?*

La hundo mordis la knabon. (subject – predicate – object) – *The dog bit the boy.*

Kiu mordis la knabon? – *Who bit the boy?*

La knabon mordis la hundo. (O – P – S.) – *The boy was bit by the dog.*

Kion faris la hundo al la knabo? – What did the dog do to the boy?

La hundo la knabon mordis. (P – S – O) – *It was biting, what the dog did to the boy.*

There are some limitation of this freedom – prepositions have to stand before its noun, adverb has to precede the word it modifies, etc.[29]

[27] Topic focus articulation is, simply said, used to distinguish between given and new, between psychological subject and psychological predicate. See Sgall, Hajičová, Panevová: The Meaning of the Sentence in its Semantic and Pragmatic Aspects, Academia, Praha 1986.

[28] However even in English, certain suggestions of “free word order” can be found: *He moved from Boston to Chicago. (Where did he move from Boston?)* vs. *He moved to Chicago from Boston. (From where did he move to Chicago?); We came to Paris yesterday. (When did we come to Paris?)* vs. *Yesterday, we came to Paris. (Where did we come yesterday?)*

[29] For more information see PAG §272-274.

2.14.4 Question

The yes/no questions are formed from indicative sentences by using particle *Ĉu* at the beginning of the sentence:

***Ĉu** vi havas domon? – Do you have a house?*

The question-word question are created by using an interrogative correlative:

***Kion** vi faras? – What do you do?*
Kiu estas tiu? – Who is that?
Kiam vi alvenas? – When do you arrive?

2.14.5 Negation

For negation of the sentence, particle *ne* before the finite verb is used.

Mi ne volas tion. – I do not want it.

It is also possible to use negative correlatives:

Mi vidas neniun. – I see nobody.

Two negations create positive statement.

Mi ne vidas neniun. – I see somebody.

3 Word building

Esperanto is a language with very rich word building. There is a large system of affixes. In addition, there are no phonologic alternations (compare *kantas – kantis – kantinta – kanto* with *sing – sang – sung – song*.[30]).

Elements of Esperanto can be classified into these categories:

1) Roots (*radikoj*):

patr – man, bon – good, ir – go

2) Affixes (*afiksoj*) – subset of roots with some specificity (see chapter 3.2)

ej – place, ism – a movement, iĝ – to become

3) Inflectional affixes or endings (*finaĵoj*) (described in chapter 2):

a) category endings: *o* – noun, *a* – adjective, *e* – adverb, *i* – infinitive verb

b) declensional endings: *j* – plural, *n* – accusative

c) conjugative endings: *a* – present, *i* – past, *o* – future, *s* – indicative, *nt* – active participle, *t* – passive participle, *us* – conditional, *u* – volitive

4) Primitive words (*vortetoj*) – a subset of roots that do not require any category ending to form a word. However, the ending is possible.

tro – too, too many, tri – three, vi – you, aŭ – or

fi – fie → fia – disgusting, anstataŭ – instead of, anstataŭi – to substitute

All Esperanto roots (without primitive words) can be classified into three categories – nominal roots, adjectival roots and verbal roots. These categories are inherent to them:

- nominal roots: *hom – human, martel – hammer, buter – butter, domo – house*
- adjectival roots: *bel – nice, bon – good, blu – blue*
- verbal roots: *kur – run, kapt – catch, dir – say*

Suffixoids[31] can be classified into these three categories too:

- nominal roots: *ul – person, ej – place, il – tool, in – feminine*
- adjectival roots: *ebl – able, em – having tendency, end – necessary*
- verbal roots: *ig – to cause, iĝ – to become*

The primitive words have different POS:

mi – I (pronoun), *tiu – this* (pron.), *apenaŭ – scarcely* (adverb), *tre – very* (adv.), *kial – why, tra – through* (preposition, *sed – but* (conjunction), *ĉu – whether* (particle)

Sometimes the category of the root is obvious, sometimes it is arbitrary set (*komb – comb* (v), *bros – brush* (n)).

If a category ending (*o* for a noun, *a* for an adjective and *i* for a verb) is added to the root of the same category, it does not change the meaning of it. In this case, the grammatical endings are redundant.

Theoretically, any root can be converted to any category just by assigning the ending of that category. The meaning of the result is depending on the category of the root:

1) Noun ending (*o*)

a) With adjectival roots – the name of the quality expressed by the root.

bela – beautiful → belo – beauty, beautifulness

b) With verbal roots – the name of the process expressed by the root.

kuri – to run → kuro – a run

2) Adjectival ending (*a*) with noun or verbal roots – the quality of or relation to the concept expressed by the root.

reĝo – king → reĝa – royal

ami – to love → ama – amatory

3) Verbal ending (*i*) with nominal or adjectival roots – action or state characterized by the concept expressed by the root.

martelo – hammer → marteli – to hammer, to work with the hammer

bela – nice → beli – to look nice

---

[30] Because of the origin of the words, it is possible to look at some Esperanto words as having also some phonological or orthographical changes: *agi, akto, reakcio – to act, action, reaction; inteligenta, intelekto – intelligent, intellect;* etc. These forms are regarded as distinct roots. See also pseudoaffixes (3.2.5)

[31] See chapter 3.2

If the meaning cannot be expressed by endings, it is necessary to use suffixes. For example, the name of the process can be expressed simply by using the noun ending only with verbal root (*kuri* → *kuro*). This is impossible with roots of another category (*marteli – to hammer*, *martelo – a hammer, not working with hammer*). In this case, it is possible to use suffix *ado* (*martelado – working with hammer*). Using this suffix followed by a noun ending after a verbal root is a redundancy, but it is used to stress the fact of the process. The quality can be expressed by the suffix *eco* (*marteleco – the quality of being a hammer*, *beleco* = *belo*). For more suffixes see chapters 3.2.1 and 3.2.2.

In the following text, the roots and affixes are very often showed with their category endings. The phrase "root *domo*" is an abbreviation for "nominal root *dom*."

As an example of word derivation I quote some words derived from the root *labori – to work*[32]:

laboro – work (n.)
labora – work (adj.)
labore – by a work
labor| isto – worker
labor| ist| ino – female worker
labor| ist| aro – labor (workers considered as a group)
labor| ego – grand work
labor| aĵo – the thing concerned with a work
labor| ebla – workable
labor| ejo – workshop, workplace
ek| labori – to start to work
labor| estro – the chief of the work
labor| ema – laborious
mal| labor| ema – lazy
labor| em| ulo – hard worker
labor| en| da – that has to be done
fi| laboro – disgusting work
labor| ilo – a tool for a work
re| labori – to do again, to start work again
labor| ulo – worker
sen| labor| ulo – unemployed person
labor| tago – work day
tag| laboro – the work for the day

3.1 Composites

The composites in Esperanto are formed by determination, juxtaposition and coordination. Nearly everything can be part of a composite (classical roots, affixes, prepositions, interjections, primitive pronouns and adverbs, numerals). Composites can have two or more parts. I will deal mostly with composites of two roots – the composites of more roots can be viewed as incrementally built words combining in each step two parts.

3.1.1 Determination

In composites of this type, one element determines the other element . Mostly, the second element is modified by the first one. The only exception are true suffixes (see 3.2 and 3.2.1) – they modify the preceding root.

Except the direction of determination, the composites can be classified also according to the result of the composition:

1) The semantic meaning of one element is modified, qualified or restricted by the other element.

2) The result is not restricted by meaning of the main element, but has a new meaning – the new meaning is a mutation of the meaning of the main element and its determiner.

duon|horo is not an hour (*horo*) that has a half (*duono*) length, but it is a *half of an hour*

Some words can be put in both groups, depending on their analysis:

antaŭ|ĝardena:

[32] Věra Barandovská: Esperanto pro samouky, p. 183 and Karel Kraft: Esperantsko-český slovník.

1) relating to something in the front of (*antaŭ*) the garden (*ĝardeno*)
2) relating to the garden in the front of the house (*antaŭgardeno*)

The inherent category of the main element is inherited by the result of the composition: *dormi + ĉambro* → *dormoĉambro*)

In the theory presented in PAG, the composition is driven by so called rules of vortefiko rekta (Direct effect of the word). It states that the inherent category of the main element has an effect on the category of the determining word (in composites the category is not expressed by an ending):

1) The nominal element makes the preceding root nominal.
2) The adjectival element makes the preceding root nominal.
3) The verbal root makes the preceding root adverbial or adjectival (with meaning of predicative adjective).

However, some words have to be analyzed by rule of vortefiko inversa (Inverted effect of the word) – the category of the main element is influenced by the determining element.

I do not know if there are any rules for distinguishing which of these rules should be used. The problem would require further study and I do not use these rules.

3.1.2 Coordination

Coordination is a composition of two or more words on the same level. It is impossible to say that one root is modified by the other.

E.g.: *blua-blanka – blue-white, angla-franca – English-French, trafe-maltrafe – better or worse*

These resulting forms can be treated as two words and can be separately declined:

*La fotoj estas nigra**j**-blanka**j**. – The snaps are black-and-white.*

Or can be treated as one word:

*La fotoj estas nigra-blanka**j**.*

Coordinating composites are mostly written as two (or more) separate words with hyphen between them. However, sometimes the ending is used only with the last element (*nordoriento – northeast*). Moreover, sometimes it is possible to see a form with endings in the middle of the word without any hyphen: (*nigra* (*black*) + *blanka* (*white*) → *nigr**a**blanka* = *nigra-blanka – black and white*). See also 3.3.2 3)

3.2 Affixes

Specific group of roots can be called affixes. These roots are mostly used with some other roots in composites. However, they can form words also alone, just by adding an ending. This case is not so common as using them in composites and not all theoretically possible forms of using affixes as roots can be found in a real text.

Affixes can be classified into two groups: affixoids (more like a classical root) and true affixes.

The difference between real suffixes and classical roots together with suffixoids is following – when two roots are put together to form a composite, the first root modifies (is a determiner) the meaning of the second (main) root:

dormo|ĉambro – room for sleeping – ĉambro – room is modified by *dormi – to sleep* (o after the root dorm is inserted for better pronunciation – see chapter 3.3)

In the case of suffixes, the determination is done in opposite direction:

dorm|egi – to sleep deeply – dormi – to sleep is modified by the word *ega – big*.

The distinction between Esperanto prefixes and roots is not so obvious. The prefixes are always determining the following root. The result of such a determination can be a modification:

dis|iri – to go in different direction – the verb *iri – go* is modified by prefix *dis.*

Or a totally different meaning:

mal|bona – bad, the meaning of the word *bona – good* is negated by the prefix *mal*.

Another difference is that at least about some true suffixes (*aĉ*, *eg*, *et* and *um*), it is possible to say that they do not have inherent category and are transparent according to the category of the stem they are assigned to.

The distinction between classical roots and true affixes is not clear and depends mostly on a tradition. Different theories put the frontier between roots (and affixoids) and true affixes in a different point. PAG uses following distinction: *bo*, *eks*, *ge*, *mal* and *pra* are called true prefixes and *aĉ*, *eg*, *et*, *um, nj and ĉj* are called true suffixes.

For my application, the division between classical roots, affixoids and affixes is not important. It will be driven mostly by practical needs – ffixes are more often used in word building

then the rest of roots, so it is good to spend more time with them. A mistake made in rules for an affixoid would produce more errors when analyzing a real text, than a mistake made in rules for an ordinary root.

I will put apart suffixes (*aĉ, eg, et* and *um*), because their transparency to the inherent category of the stem and because of different direction of modification of the meaning in composites. I will make no distinction between prefixoids and prefixes.

There are two purposes of this chapter. The first purpose is to prepare direct background for the implementation (prefixes *pra* and *bo*, suffixes *io*, *ujo*, etc). The second purpose is to show how flexible is the Esperanto word building, to show that it is hard to say that something is impossible. For this reason, the paragraph "Used as a root:" is added. Therefore some important affixes are described in separate chapter, to some is devoted only few lines.

3.2.1 True suffixes

3.2.1.1 Aĉ

The suffix *aĉ* gives to the stem shading of contempt or disgrace, detestation.

ĉevalo – horse → ĉevalaĉo – nag
domo – house → domaĉo – hove
hundo – dog → hundaĉo – cur
paroli – speak→ parolaĉi – tittle-tattle, twaddle

See also and *mis-* in chapter 3.2.3.5.

Used as a root:

aĉa – useless, ugly, aĉaĵ(ar)o – junk, lumber, aĉigi – disgust, make something terrible, aĉulo – ugly fellow, aĉularo – mob, rabble

3.2.1.2 Eg

The suffix *eg* augments or strengthens the idea shown by the root.

urbo – town → urbego – big town, city
domo – house → domego – big house, mansion
varma – hot → varmega – very hot, boiling hot
necesa – necessary → necesega – absolutely necessary
tre – very → treege – extremely
ridi – laugh → ridegi – cachinnate, guffaw

Used as a root:

ega – enormous

3.2.1.3 Et

The suffix *et* is used to form diminutives.

urbo – town → urbeto – small town
domo – house → dometo – big house, cottage
varma – hot → varmeto – warm

This suffix can be also used with the names of persons (or family members) to make intimate forms (see also suffixes *ĉj*, *nj*: 3.2.2.5)

Paŭlo – Paul → Paŭleto – Paul, my bonnie
patro – father → patreto – dad

Used as a root:

eta – tiny, etulo – small child, etulino – small girl, etaĵo – small thing, etigi – to diminish

3.2.1.4 Um

The suffix *um* has not defined meaning. It just somehow modifies the meaning of the root. The meaning of the resulting word is hard to decode from the knowledge of the root – the only definite thing is that, they have something in common.

vento – wind → ventumi – to ventilate
kolo – neck → kolumo – collar

akvo – water → *akvumi – to water, to irrigate, to sprinkle water on*
plena – full → *plenumi – to fulfill*

Used as a root:

umo – doohickey, umi – to do something (If you can not find the right word)

3.2.2 Suffixoids

3.2.2.1 Igi

Verbs created by the suffix *ig* mean: *"to cause, to do something, to be in the state of the stem"*. Intransitive verbs are changed into transitive. Verbs with this suffix are often told to be in factitive voice. The suffix *ig* is very often.

blanka – white → *blankigi – to make something white, whitewash*
dormi – sleep → *dormigi – put to sleep*

If the suffix is followed by the noun ending the result is the name of the action (*pura – clean, purigi – to clean, purigo – cleaning*). If the suffix is followed by an adjective ending the result is adjective with meaning doing, able to do or relating to something. (*puriga – purifying, able to clean, cleaning*).

The verb *igi* means to cause.

3.2.2.2 Iĝi

Verbs created by the suffix *iĝ* mean: *"to become, to turn into"*. Transitive verbs are changed into intransitive. Verbs with this suffix are often told to be in mediopassive voice. The suffix *iĝ* is also very often.

ruĝa – red → *ruĝiĝi – to become red*
naski – to born → *naskiĝi – to be born*

It is possible to form add noun ending (name of the action). Adjectival and adverbial endings are possible, but latter not very often.

The verb *iĝi* means to become.

3.2.2.3 Ado

The suffix *ad* emphasizes the process. With the noun ending, it means the name of the action, with the verbal ending, it means repetition or long lasting of the process. For verbal roots, it has the same meaning as adding simple noun ending.

martelo – hammer → *martelado – hammering, marteladi – use hammer often*
iri – go → *ir(ad)o = the act of going, iradi – to be going for some time*

Used as a root:

ada – continual

3.2.2.4 Eco

Suffix *ec* means quality. The best English counterpart is a suffix *-ness*. For adjectival roots, it has the same meaning as adding simple noun ending.

riĉa – rich → *riĉ(ec)o – richness*
konfuzita – confused → *konfuziteco – confusedness*

Used as a root:

eco – quality, characteristic, ece – in a characteristic way, eca – having the character, ecaro = karaktero – character (sum of qualities)

3.2.2.5 Ĉj, Nj

The suffixes *ĉj* (for males) and *nj* (for females) make from the root an intimate form. The root can be shortened – the suffix is attached after one of the first five letters.

Johanno – Jack → *Joĉjo – Jack*
Johana – Joan → *Jonjo – Joanie*
patro – father → *paĉjo – dad, panjo – mum*
filo – son → *fiĉjo – little son, finjo – little daughter*
frato – brother → *fraĉjo – little brother, franjo – little sister*
amiko – friend → *amiĉjo – dear friend* (he), *aminjo – dear friend* (she)

These suffixes are now not very often used (except few words like *paĉjo*, *panjo*) and are very often replaced by suffix *et* (see 3.2.1.3) or by national intimate forms (*Johnny*, *Dick*, *Saŝa*).

3.2.2.6 Other suffixoids

Aĵo

The suffix *aĵo* forms a concrete, perceivable manifestation of the root.

nova – new → novaĵo – new thing, novelty
fotografi – to take photographs → fotografaĵo – photography

Used as a root:

aĵo – a think

Ano

The suffix *ano* forms a member, participant, resident.

klubo – club → klubano – member of the club
Budho – Buddha → budhano – Buddhist (see also *isto*)

Used as a root:

ano – a member of a club, society, ani – to be a member, anigi – to make somebody a member of something, aniĝi – to become a member, aneco – membership, aniĝilo – application form, etc.

Aro

The suffix *aro* adds to the root the meaning of a collection.

arbo – tree → arbaro – forest
homo – man → homaro – mankind (≠ *aro da homoj – group of people*)

Used as a root:

aro – group, ara (adj.), *are* (adv.) *– in groups, grandare – in big groups, ari – to be in group, ariĝi – to group oneself, arigi – to group somebody*

Ejo

The suffix *ejo* means place where something is performed or where something is kept, a building, etc.

lerni – to learn → lernejo – place for learning, school
preĝi – to pray → preĝejo – a church, musk, etc (*kirko = church*)
ministro – minister → ministrejo – department

Used as a root:

ejo – place

Ero

The suffix *ero* means an element of the thing expressed by the root. Not nominal stems are automatically nominalized (as if suffix *aĵo* were added).

pano – bread → panero – a crumb of a bread
neĝo – snow → neĝero – snow flake
kudri – to sew → kudrero – a stitch

Used as a root:

ero – element, grain, ereto – small element, grandera – coarse-grained, diseriĝi – to disintegrate (intransitive), *diserigi – to cause to disintegrate*

Estro

The boss of the thing expressed by the stem.

ŝipo – ship → ŝipestro – captain
urbo – town → urbestro – mayor

The boss of some group of people can be expressed also by prefixing the root *ĉefo – chief.*

Used as a root:

estro – boss, estri – to direct, estraro – board of directors

Ido

The suffix *id* forms an offspring, young creature, etc.

hundo – dog → hundido – puppy
planto – plant → plantido – small plant

latina – Latin → latinida lingvoj – languages with the Latin origin

Used as a root:

ido – offspring, ideto – small young, idaro – all descendants

Ilo

The suffix *il* means a tool for doing whatever is expressed by the preceding root.

trancĉi – to cut → trancĉilo – tool for cutting, a knife
komputi – to compute → computilo – a computer
butero – butter → buteri – to butter → buterilo – knife for buttering
linio – line → linii – to line → liniilo → ruler

Used as a root:

ilo – instrument, ilaro – set of tools, ilujo – box for tools, ilejo – workshop

Ino

The suffix *in* means a female.

patro – father → patrino – mother
bovo – cow → bovino – she-cow

Esperanto is a sexist language. Most of roots with meaning of human beings are of the male sex. Today more and more of them are considered neutral. Especially professional titles are neutral. The male equivalent of this suffix is a prefix *vir-*, it can be added to roots which sex is neutral.

kato – cat → virkatino – tomcat

Ingo

The suffix *ingo* has a meaning of a holder for the thing described by root.

kandelo – candle → kandelingo – candle-holder

Used as a root:

ingo – holder, mostly sheath, ingi = eningig – to put into the holder, sheath, malingi = elingigi – to put out of the holder, unsheathe

Ismo

The suffix *ismo* has a meaning of a doctrine, movement, system, etc.

Budho → budhismo, Markso → marksismo
ĵurnalo – newspaper → ĵurnalismo – journalism

Used as a root: *ism – movement, doctrine*

isto

The suffix *isto* has a meaning of an individual professionally occupied with something, somebody who is used to do something. The suffix can be also used as equivalent for two suffixes *ist|ano*. This is not used if the *ismo* is added to the member of the movement, etc. (*krist|ano – Christian → kristanismo – Christianity*).

labori – to work → laboristo – a worker
lingvo – language → lingvisto – linguist
Budho → budhisto, Markso → marksisto

Ujo

The suffix *ujo* has three meanings:

1) A container or box for something. This is the main meaning.

papero – paper – paperujo = paperkesto – box for paper
salo – salt → salujo – saltcellar, saltshaker

2) A tree having fruit or flowers specified by stem. This meaning of the suffix is rather archaic, today it is replaced by forming a composite with *arbo – tree* or *arbeto – small tree*.

pomo – apple → pomujo = pomarbo – apple-tree

3) A country for the nationality expressed by the stem. See chapter 3.3.4 Names of countries.

Italo – Italian → Italujo = Italio – Italy

When the suffix *ujo* is used as a root, it has the first meaning – a box or container.

Ebla

The meaning of suffix *ebla* is "*suitable for being done*".

legi – to read → legebla – readable
fari – to do → farebla → possible to be done

Used as a root:

ebla – possible, possible to be done, eble – maybe, eblo = ebleco – possibility, eblaĵo – possible thing, possibility, ebligi – enable, ebliĝi – to become possible, malebla – impossible

Ema

The meaning of the suffix *ema* is "*to have tendency or inclination to do the thing described by the stem.*"

labori – to work → laborema – industrious
dormi – to sleep → dormema – sleepy

Used as a root:

ema – inclining, emo – inclination, emi – incline, emiĝi – to become inclining, emigi – to cause that something is inclining

Enda

The meaning of the suffix *enda* is "*it must be done the thing described by the stem*"

skribi – to write → skribenda – that must be written
vidi – to see → videnda – that must be seen

Used as a root:

enda – mandatory, endo – necessity, endi – it is necessary

Inda

The suffix *inda* has meaning "*worthy -ing*"

fari – to do → farinda – worth doing
vidi – to see → vidinda – worth seeing

Used as a root:

inda – worthy, indi – to be worthy, indigi – to make something worthy, malinda – to be unworthy, senidulo – unworthy man

3.2.3 Prefixes

3.2.3.1 Bo

Bo marks relative by marriage. In English, the same thing is done by adding *in-law*.

bofrato – brother-in-law, bofilo – son-in-law, bopatrino – mother-in-law, bokuzo – cousin-in-law, bonevo – grandson-in-law, etc.

The exception is a word for child coming from the previous marriage(s) of one of the spouse – it is marked by prefixing *duon*[33].

Bo can be also used in following words:

boparenco – relatives by marriage
bofamiliano – member of the family by marriage
boedziĝi – to marry with the wife of one's dead brother (used by Zamenhof in the translation of Bible)
boamiko – jocular way to call friend of one's spouse

There is no distinction between relatives got by one marriage and relatives got by two marriages: *mia bofrato* is brother of my spouse or husband of the sister of my spouse.

If *bo* is together with prefix *ge* (see 3.2.3.2), *bo* stands before *ge*[34]: *bogefratoj – brothers-in-law and sisters-in-law.*

The prefix as a root:

boulo = boparenco – relative-in-law, boeco = boparenteco – the type relation between two relatives-in-law, boa – being of the in-law type of relation

[33] See chapter 2.8.2.6 - Fractions.

[34] This is opinion of the PAG (§417). PMEG (*bo*) is not so strict, it states that the order of the *bo* and *ge* does not matter, and that it is only a habit to put *ge* first.

3.2.3.2 Ge

Ge marks both sexes. This prefix is used before roots of male beings or roots that are neutral from the point of the sex.

geknaboj – boys and girls, gejunuloj – young people, youth, geinstruistoj – teachers of the both sexes

For neutral words, the prefix is very often not necessary. Words like *lernantoj* (*pupils*) or *doktoroj* (*doctors.*) are good enough for describing beings of both sexes and the prefix *ge* is used only for stressing the fact that people in the group are of both sexes.

The plural after the word with this preposition is common, but not necessary: *geedzo – spouse, gepatro – parent.*

The meaning of the prefix can be slightly different, depending on context:

1) A pair (e.g. of husband and wife): *gepatroj – mother and father, parents, geonkloj – uncle and aunt, geedzoj – husband and wife, geamantoj – lovers.*

2) Members of the same type, but both sexes of a family*: gefiloj – sons and daughters of the same family.*

3) The whole family: *geurbestroj – the family of the mayor*

Prefix *ge* is also used with things that do not have sex. In that case, it means that the things are related with both sexes: *gelernando – coeducation, gelernejo – coeducational school*. However, this usage is quite rare.

As root:

geo – he and she, pair, gea – mixed, etc. (*gea lernejo = gelernejo – coeducational school*), *geiĝi = pariĝi – make pairs from oneself, geigi = parigi – to pair*

3.2.3.3 Mal

Prefix *mal* denotes total opposite to the stem.

malbela – ugly, malvarma – cold, malgranda – small, malrapida – slow, malami – hate, malaperi – disappear, malamiko – enemy, malantaŭ – behind.

Prefix mal is very often used. In the beginnings of the language, it was nearly the only way of finding an opposite for most of the words. Even very common words (*malgranda – small, malfermi – open, malnova – old,* etc) had to be expressed using this prefix. Today, some synonyms to the mal-words exist, some of them are used more and some of them less often: *fini = malkomenci – finish, frida = malvarma – cold, breva = mallonga – short, eta = malgranda – small, dura = malmola – hard, cis = maltrans = on this side* (*trans – across*), *olda = malnova – old,* etc.

As a root:

malo – opposite (noun), *mala – opposite* (adj.) *male – opposite* (adv.), *oppositely, malinda – undesirable, unwelcome,* etc.

3.2.3.4 Pra

Prefix *pra* has following meanings:

1) With names of relatives, one generation older or younger: *praavo – great-grandfather, pranepo – great-grandson, praonklo – great-uncle.* The prefix *pra* can be even repeated: *prapraavo – great-great-grandfather*. For the father of the father and for the son of the son are words *avo* and *nepo – prapatro* and *prafilo* belong to the second category.

2) Very distant in time (mostly in the past – ancient or primeval): *praarbaro – primeval forest, pratempo – primeval ages, prahistorio – prehistory, prabesto – primeval animal, prahomo – primeval human, prapatro – founder of the family, of the kin or nation, prafilo – descendants after many generations.*

As a root:

prae – primevally, praa – primeval, praeco – "primevalness", praulo = primeval ancestor, forefather

3.2.3.5 Other prefixes

Eks

Eks marks something former. It is mostly used in front of the word with the meaning of some profession or function.

***eks**prezidanto – ex-president, **eks**posedanto – former owner, **eks**direktoro – former director, **eks**urbestro – former mayor, **eks**edzo – former husband, **eks**edziĝi – to divorce oneself, **eks**moda – out of the fashion*

Eks used as a root: *eksigi – to force somebody to abdicate, eksiĝi – abdicate, leave a club, eksa – quondam, abdicated, eks*! – *Eks pri la reĝo! – Away with the king!*

Dis

Prefix *dis* means separation in different directions, scattering.

iri – to go → disiri – to go in different directions
vojo – way → disvojiĝo – road-fork

Ek

Prefix *ek* means the beginning or ephemerality.

iri – to go → ekiri – to start to go, to set out
krii – to shout → ekkrii – to shout out

Ek can be also used alone as an interjection: *Ek! – Let's start!, Ek al li!*[M] *– Let's catch him.* Or as a verb: *eki* = *komenciĝi – start* (intransitive), *begin, ekigi – start* (transitive)

For

Prefix *for* means removing, disappearing, distance, spoiling.

iri – to go → foriri – to leave
dormi – to sleep → fordormi – to spoil time by sleeping

Mis

Prefix *mis* expresses an error or incorrectness. In contrary to the suffix *aĉ,* this prefix is used in objective stating.

kalkuli – calculate → miskalkuli – miscalculate
traduki – translate → mistraduki – mistranslate
kompreni – understand → miskompreni – misunderstand

Mis used as a root:

misa – incorrect, erroneous, misi – to err

Re

Prefix *re* means returning or repetition.

veni – to come → reveni – to come back, to return (intransitive)
meti – to place → remeti – to put back, to return (transitive)
legi – read → relegi – read again

Re can be used alone in *ree – again, reen – back, rea* – adjective from *ree* or *reen*.

3.2.4 Unofficial affixes

There are also many unofficial affixes. I list only the most often ones:

-iva – capable of doing something
produkti – product → produktiva – productive
-eska – similar to, or *in the manner of*
japano – a Japanese → japaneska – Japanesque
-ala – is used to derive adjectives from nouns derived from adjectives
varma – hot → varmo – heat → varmala – thermal
-oida – resembling; having the appearance of; related to, mostly technical
antropo – human → antropoido – anthropoid
-oza – full of
poro – pore → poroza – porous
-iza – to apply something (thing or method) to an object
salo – salt → salizi ion – add salt to something
retro- – in the opposite direction
iri – go → retroiri – to go in opposite direction

There is also large amount of affixes used in some special field – in chemistry (*-oza – feroza – ferous, -ika – sulfika – sulfuric,* etc), in botany, medicine (*-ozo – sklerozo – sclerosis, -ito – dermatito – dermatitis*) and so on.

3.2.5 Pseudoaffixes

Some of the unofficial affixes are partly so called pseudoaffixes. They are mostly affixes in the languages the Esperanto vocabulary comes from.

Many of Esperanto roots are composites in the language they come from. Therefore, some roots start or finish with the same sequence of characters. These sequences look as an affix. However, the rest of such a word is very often not an Esperanto word. On the other hand, sometime new words are created connecting these sequences with Esperanto roots. These elements are called pseudoaffixes (*pseŭdoafiksoj*).

Typical example is a pseudosuffix *logio*: *ornitologio – ornithology, zoologio – zoology*, etc. However, there are also words *metodologio – methodology* (*metodo – method*), *antropologio – anthropology* (*antropo – human*). And there are also purely Esperanto words: *esperantologio – science about Esperanto, birdologio – ornithology, formologio – morphology*, etc. In these words is a suffix *ologio*.

Another pseudosuffixes are *iko* (*poeto – poet → poetiko – poetics, stilisto – stylist → stilistiko – stylistics*), *acio, icio* (*delegacio – delegation, operacio – operation, pozicio – position, etc*) and many others.

There are also pseudoprefixes: *aŭto* (*autobiografio – autobiography, aŭtomobilo – car, aŭtonomio – autonomy, aŭtokrato – autocrat*), *anti* (*antikristo, antisemito*), *eŭ* (*eŭfemismo – euphemism, eŭgeniko – eugenics*) and others.

Most of these words are treated as separate roots in Esperanto. Some of them can be considered as unofficial affixes (*ologio, iko*). In contrary to classical affixes, these cannot be used systematically (sometimes is ok *ologio*, sometimes *iko*).

3.3 The rest

3.3.1 Inserted o

To make the pronunciation easier, it is possible to insert a vowel *o* between two roots in a composite: *puŝ**o**ŝipo – tugboat, skrib**o**tablo – writing desk.* With some words the letter o is inserted to make them more recognizable, because of the tradition or because the words have ends with an o in international usage: *di**o**similа – like a god, radi**o**elsendi – radiobroadcast.*

The letter *o* cannot be inserted in front of suffixes or after prefixes.

In PAG[35], this problem is described differently – as conserving of endings. The conserving of endings has the same reason as inserting of an *o*. However, in this case the o between two roots is not an inserted vowel but an nominal ending of the first root. Another consequence is, that also adjectival ending a and adverbial e can be found between two roots. The type of the ending is driven by so called vortefiko rules (vortefiko – effect of the word)[36].

This theory seams reasonable to me. Except the problems with vortefiko rules, I do not know if the theory could be confronted with real data. This problem would require further study, especially of a large corpus. Provisionally, I look at the *o* as an inserted character.

3.3.2 Hyphen

According to PAG[37], hyphen (*dividstreko*) in composites is used in following cases:

1. In composites with three or more roots (not counting affixes) to show theoretical bracketing of the main and determining elements: *vapoŝip-asocio – steamboat association in contrast to vapor-ŝipasocio – steamy boat-association*
2. In composites with two roots, to make them more easily to recognize: *sen-tema – without any theme;* especially when the second root starts with a vowel: *bel-aspekta – looking pretty*. The hyphen is not used before suffixes (incl. suffixoids) and after prefixes (incl. prefixoids). However, it is recommended to use it, if the affixoid is used as classical root: *il-riparo – reparation of the tool* (*ilo – a tool*, a suffixoid).

[35] See PAG §309
[36] See chapter 3.1.1.
[37] See PAG §14B – I.

3. It is recommended to use it in coordinative composites (see 3.1.2): *membro-abonanto* – *member-subscriber*; and it is necessary to use it if the elements are inflected: *esperantistoj-amikoj* – *Esperantists-friends.*

For other uses of a hyphen, see also chapters 2.4.1.1 Declination of proper names and 3.3.5 Abbreviations

3.3.3 Sciences

Names of sciences are full of pseudosuffixes[38], however some of the sciences can be regarded as composites.

The most often suffix is *io* (unofficial), that makes the name of the science from the scientist. The scientist ends very often with *ologo* – partly pseudosuffix (*astrologo*, *ekologo*), partly unofficial suffix (*antrop|ologo*, *soci|ologo*).

Words for sciences sometimes contain pseudosuffix *iko* (*poetiko* – *poetics*, *stylistko* – *stylistics*). *Ik* can be regarded in some cases as suffix forming the name for the science from the scientist (*stylisto* – *stylist*). However these cases are very rare and the rest before the *iko* is very often not a scientist (*simbolo* – *symbol*, *simboliko* – *symbolism*) or the result is not a science (*gimnasto* – *gymnast*, *gimnastiko* – *gymnastics*).

3.3.4 Names of countries

The problem of the names of countries and nationalities has been often discussed. There are two ways – to form the name of the inhabitant from its country or vice versa. The current state of the names of countries is evolved tradition, international influences and tendency to use some simple system.

Originally, an inhabitant was primary for the Old Word and a country for the New World; with some exceptions. The inhabitant was formed by the suffix *ano* and the country by suffix *ujo*. Names of some were derived from a town or a river by the suffix *io* (*Meksiko* – *Ciudad de Mexico* → *Meksikio* – *Mexico*).

However, there was a tendency to make the names more international. Some names were using the word *lando* (*Finnlando*), the suffix *io* was used more and more instead of the suffix *ujo* and a new suffix *istan* was used for some countries.

Today, there is a list of standard names of countries (Listo de normaj landnomoj)[39]. This list put all countries into two categories and some subcategories:

1) Country is primary, inhabitant is formed by suffix *ano*.
   Peruo → *Peru|ano*, *Aŭstralio* → *Aŭstrali|ano*, *Nepalo* → *Nepal|ano*
2) Inhabitant is primary, country is formed by various suffixes:
   a) by the suffixes *io* or *ujo*.
      Hungaro → *Hungario*/*Hungarujo*, *Turko* → *Turkio*/*Turkujo*
   b) by the root *lando*.
      Finno → *Finnlando*, *Skoto* → *Skotlando*
   c) by the suffix *istano*.[40]
      Uzbeko → *Uzbekistano*, *Afgano* → *Afganistano*

The names derived from the name of a town or a river by the suffix *io* are in the first category.

3.3.5 Abbreviations

Abbreviations (mallongigoj) have nearly the same form as in other languages:

E.g.: *ekz.* – *ekzemle* – *for example*, *k.t.p.* – *kaj tiel plu* – *etc.*, *p.* – *pago* – *page*, *t.e.* – *tio estas* – *i.e.*, *PIV* – *Plena Ilustrita Vortaro* – *The Full Illustrated Dictionary*

---

[38] Sequences of characters that very often repeat in Esperanto words, mostly suffixes in languages the words originate from. See chapter 3.2.5.

[39] Oficialaj Informoj de la Akademio de Esperanto, n-ro 9, 1989

[40] *Pakistano* belongs to the lexicon country, the inhabitant is called *Pakistan|ano*.

Very often, the abbreviations are formed by conserving few letters from the beginning and possibly some from the end and by replacing the rest by a hyphen. Such abbreviation has grammatical ending and is normally declined.

d-ro – doktoro – doctor, s-ro – sinjoro – Mister, s-rino – sinjorino – Mistress

4 Implementation

4.1 Two-level morphology

Two level morphology was first presented by Kimmo Koskenniemi, a Finnish computer scientist, in his dissertation[41]. System using two level morphology has two main parts – linked lexicons and two-level rules. The basic idea is that lexicons contain morphemes and the links between lexicons specify the possible cooccurrences and relative order of morphemes. Two-level rules are used to transform morphemes to the surface level (to the orthographical or phonological representation) or back. There must be a bijection between symbols on both levels. Each rule can be expressed by a finite state automaton. All automata for rules are then compiled into one big finite state automaton.

For example, the English present participles are expressed by adding the suffix *-ing* to the verb. Therefore, the lexicon of verbs would contain a link to the lexicon containing the suffix *-ing* (and maybe *-ed*): *wait* + *ing* → *waiting*. However in the participle *writing* the suffix *-ing* causes the loss of the final *e* of *write* – *write* and *writ* are allomorphs of the same morpheme with different distribution. The two levels of this word would have following form:

```
Lexical form:  w  r  i  t  e  |  i  n  g
Surface form:  w  r  i  t  0  0  i  n  g
```

The automaton replaces one character after another. The rules it is compiled from specify the replacement and the context in which it is possible. The context can be described using characters from both levels. Rules have to make all phonological or orthographical changes and remove all auxiliary markers. The whole set of rules is a conjunction of all single rules.

In my system I have used the program PC-Kimmo Version 2, for more information about this program see Resources.

Lexicons

Each lexical entry has four main parts: lexical form, name of the lexicon, continuation class and gloss. There are written in following format:

```
\lf |dom<¤o>=
\lx root
\alt afterRoot
\eng |house
\cze |du°m
\deu |Haus
```

There is a declaration assigning a meaning to the fields for PC-Kimmo. This is done by four commands:

```
FIELDCODE lf   U     ;lexical item
FIELDCODE lx   L     ;sublexicon
FIELDCODE alt  A     ;alternation = continuation class
FIELDCODE eng  G     ;gloss
```

It is possible to change the last line to the following:

```
FIELDCODE eng  G     ;gloss
```

That will use the field `deu` as a gloss. Fields not assigned are ignored by PC-Kimmo.

The continuation class is declared using the command `ALTERNATION:`

```
ALTERNATION afterRoot      ending suffix root
```

This means that the lexical entry using the continuation class `afterRoot` can be followed by entries from lexicons `ending`, `suffix` and `root`.

There must be a main lexicon file. This lexicon file declares all continuation classes, assigns field codes and includes files that contain lexical entries:

```
INCLUDE PIV.lex    ;file of roots
```

Formalism of two level rules

As and example I have chosen the rule stating that any *ĥ* after an *r* can be replaced by *k*:

[41] Koskenniemi, Kimmo: *Two-level morphology: A general computational model for word-form recognition and generation, 1983*

```
hx:k => r __
```

The basic struture of any two-level rule can be expressed by following schema:

```
CP op LC __ RC
```

The meaning of its parts:

1) CP – **the correspondence part** – it describes the pair of lexical and surface characters that is restricted by this rule. In my example, the digraph (treated by the system as single character) `hx` is replaced by letter `k`.

2) op – **an operator** – The operator is used to express the relation between the context and the correspondence part. There are four types of operators:

<=> the correspondence always and only occurs in the specified context
=> the correspondence only occurs in the specified context
<= the correspondence always occurs in the specified context
/<= the correspondence never occurs in the specified context

3) LC and RC – **left and right contexts** – The context defines the phonological and morphological conditions for the correspondence part.

The correspondence part and context are expressed by so called **regular pair expressions**. These expressions are very close to classical regular expressions. These expressions will be defined in next paragraphs.

Concrete pair is a pair of lexical and surface characters (including zero character – 0). This pair is expressed as `l:s`, where `l` is a lexical character and `s` is a surface character.

There are two special symbols `0` – the zero character, and # – the word boundary.

It is possible to define **sets of characters**, e.g. set of Esperanto vowels:

```
SUBSET V   a e i o u ux
```

The name of the **set of all characters** is declared by command `ANY` followed by the selected symbol:

```
ANY @
```

The **complement of the set** is expressed by preceding the set by the symbol ¬, thus `¬V` means any character except a vowel.

Abstract pair `X:Y` is a set of pairs where lexical character belongs to the set `X` and surface character belong to the set `Y`. There are also so called semiabstract pairs – `X` or `Y` in `X:Y` is a set with one element only, e.g. `@:a` – any lexical character can be represented as *a* on the surface level.

When the lexical and surface symbols in a pair are identical the correspondence can be abbreviated as the symbol alone: `a` means `a:a`, `V` means `V:V`.

The regular pair expression (RPE) can be:

1) Concrete or abstract pair, e.g. `a:c`, `a`, `V:0`, `a:V`, etc.

2) Sequence of RPEs, e.g. `a r hx:k i`

3) An alternative of RPEs: `[hx|hx:r]`

Optional parts in a sequence are written in parentheses, e.g. `C(o:0)` is equivalent to `[C | C o:0]`. Parts that can be repeated (zero to n-times) are enclosed in parentheses followed by the Kleene star: `C(X)*`.

4.2 General approach

In Esperanto, there are no phonological alternations and nearly no irregularities, therefore, there is a tendency to think that the morphological analysis must be very easy. The inflection is totally unambiguous. The problem is that the word building is very rich. There is a large set of affixes – short morphemes that are widely used. Moreover, as was shown in the chapter 3, nearly all cooccurrences of various morphemes are allowed. The only limit is the fantasy of the Esperanto speaker. Therefore, one word can be analyzed by many different ways. This is mostly no problem for a human – with the knowledge of the world and of the context. However, the context is not available on this level of linguistic representation and the knowledge of the world is a very complex problem. Some results can be obtained with classification of roots and assigning features to them. This approach was used for some simple things (prefix *pra* or *bo*), however it is very time consuming if the feature has to be assigned to a large set of roots (e.g. mass nouns).

I have tried to allow as much flexibility to the word building as possible, with some restrictions of surely impossible combinations. To achieve this goal, I have used a mixture of linking lexicons and using two-level rules with auxiliary symbols.

4.2.1 Why not generation

In my thesis, I do not care about generation. There are two reasons for it – first the set of generated forms is a subset of analyzed forms, because some words can be expressed by many ways, from which some forms are more common than other are. Very good example is the *o* inserted between morphemes to make the pronunciation easier. It is possible, but it is used only sometimes – very often depending on the nationality of the speaker. The analysis tool must allow distinguish all these possibilities, however if I would use the same system for generation, for word with three morphemes I would obtain four possibilities. The problem is with a hyphen – it can be inserted between morphemes to make the recognition easier. The number of possibilities would even increase.

Many geographical names have more than one variant (original, international, Esperanto, etc.), the ending can be assigned using a hyphen, directly with an added *o*, or with hyphen and added *o*. For example, it is possible to see all these accusatives of *Plzeň* (the town in western Bohemia): *Plzeň-on, Plzeňon, Pisen-on, Pilsenon, Pilzen-on, Pilzenon and even Pilsen-n, Plzeň-n, Pilzen-n.*

It is necessary to analyze them all, eventually with some remark about the strange, archaic or unofficial form. However, the generator should use only one of them. This requires to keep two versions of the system – one for the generation and one for the analyzes. Most of the things would be incommon, however some rules, lexical entries and continuation classes would be different. It would be hard to develop paralelly both versions, therefore I have selected to do analysis.

4.2.2 Conventions used in the following text

All features, markers and other auxiliary symbols are put at the end of the lexical entry enclosed in angle brackets. The exception is a symbol | that marks the beginning of the morpheme and = that is put at the end of a root entry to distinguish it from other morphemes[42]. All these symbols are treated by the system as one character (a multigraph) and have always zero surface realization.

In this chapter very often the word region is used. It means some part of the word building (number, names, correlatives, etc.). The regions are not strictly defined and very often they have some parts in common.

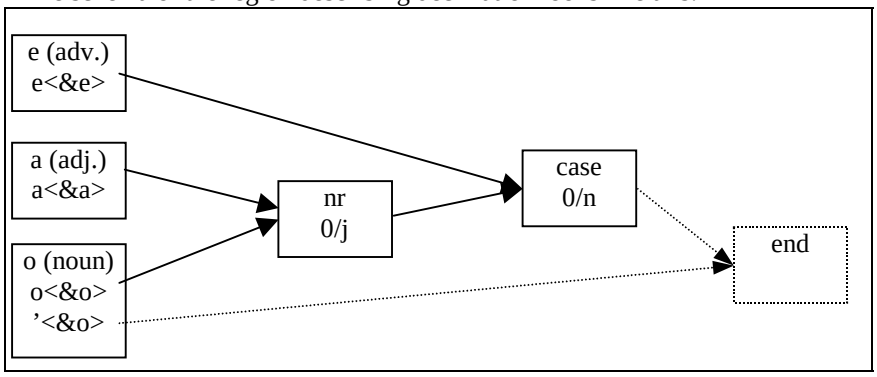
Sometimes I use a figure to show a basic structure of the region – lexicons and linkage between lexicons. The lexicons outside of the region are depicted in a dotted box. The continuation classes that are in common for all entries of the lexicon start at the rim of the box; the continuation classes that are specific for some entry start near this entry. If there are more entries with the same continuation classes, a brace is used.

4.3 Inflection

As have been said in chapter 3, the stem can easily be converted into different categories (parts of speeches) just by assigning different endings. Nominal endings are followed by the ending of the number (which is zero for singular), that is followed by the ending of the case (which is zero for nominative).

Adverbs can have accusative ending only if they are adverbs of place. Because of the complexity of the determination whether an adverb has spatial meaning or not – it would require classification of all roots in the dictionary. I will simply allow assigning the case ending to any derived adverb (the adverb with ending *e*).

The schema of the region describing declination looks like this:



[42] See chapter 4.5.1 Inserted o.

The category endings are in different lexicons to allow linking to only some of them. The same work by different means is done using their features &o, &a and &e[43]. The lexicon for the noun endings has two entries – for full form and for shortened form with apostrophe.

The lexicons have following form:

Lexicon o (noun ending):

```
\lf '
\lx o
\alt nr
\eng |xNounShort

\lf |o<&o>
\lx o
\alt nr
\eng |xNoun
```

Lexicon a (adjective ending):

```
\lf |a<&a>
\lx a
\alt nr
\eng |xAdjective
```

Lexicon e (adverb ending):

```
\lf |e<&e>
\lx e
\alt case
\eng |xAdverb
```

Lexicon nr (number):

```
\lf 0
\lx nr
\alt case
\eng
```

Because I can consider the singular as unmarked when compared to plural and because of simpler output, I will add no gloss for singular.

```
\lf |j
\lx nr
\alt case
\eng |xPlural
```

Lexicon case:

```
\lf 0
\lx case
\alt end
\eng
```

For the same reasons as by singular, I will add no gloss for nominative.

```
\lf |n
\lx case
\alt end
\eng |xAccusative
```

The continuation classes are obvious.

4.4 Verb

In this chapter I have to solve the simple verbal forms, I do not care about complex verbal forms, they are not part of morphology. The simple verbal forms can be separated into two groups –

[43] See chapter 4.6 Category prohibiting rules.

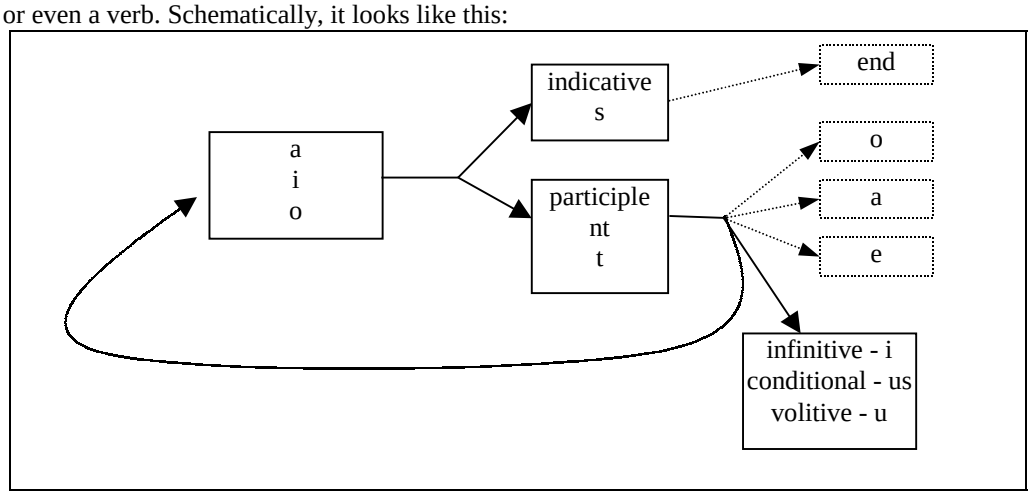
group of forms distinguishing tense (indicative and participles)[44] and a group of forms that do not distinguish tense (infinitive, conditional and volitive)[45].

The latter group is easy to handle – the endings are just assigned to the stem, and it is impossible to put anything after them (except coordinative composites[46], of course).

infinitive: *kapti|i*
conditional: *kapt|us*
volitive: *kapt|u*

That means, this group can be expressed by one lexicon containing these three endings with identical continuation classes, leading to the lexicon end.

The group of forms distinguishing tense is a bit more complicated. The indicative is formed by adding a vowel of tense[47] to the stem, followed by the ending *s*. The "participles" start the same way – a vowel of tense, then the suffix *nt* or *t* and finally the category ending for a noun, adjective, adverb or even a verb. Schematically, it looks like this:



The problem is that it is possible to add an infinitive *i* after the participle suffix and conjugate the resulting form, but it is impossible to form a participle from that. In short, it is impossible to form a participle from a participle. Therefore, I have to forbid two participle suffixes in one word.

The two level rules are the best solution. I will introduce a marker &part and add it to the participle suffix. The marker will be realized as a zero on the surface level and a rule will forbid two &part markers in one word:

```
RULE &part /<= &part @* __
```
[48]

Finally, I merge the lexicon containing the endings for infinitive, conditional and volitive with the lexicon containing the vowel of tense. They can be both added to the stem to form a verb (or a form derived from a verb). Otherwise, I would have to put into each continuation class of the stem forming a verb both lexicons.

This merging will require the items of the lexicon to have different continuation classes. The infinitive, conditional and volitive one and vowel of tense another.

The participle can be followed by different suffixes and suffixoids:

estonta – going to be → *estont**ec**o – quality of "going to be", the future* (= *futuro*)
vojaĝanto – voyage → *vojaĝant**in**o – female voyager*
mia konato – one, whom I know, my friend → *konat**ig**i – introduce*

Therefore, I add to continuation class of the participle suffix a link the lexicons of suffixoids and suffixes:

```
ALTERNATION afterPart      o a e verb suffixoid suffix
```

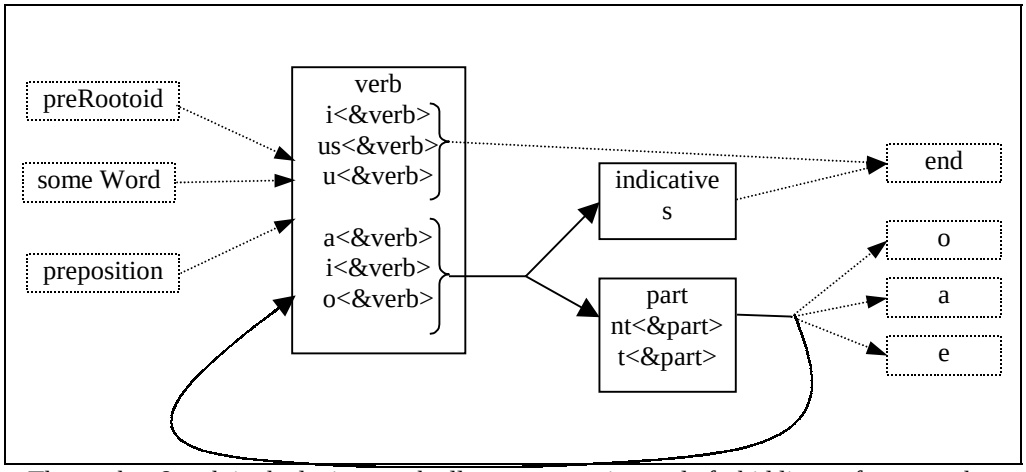
The complete schema of the region describing verbal forms looks like this:

---

[44] See chapters 2.9.3 Indicative and 2.9.6 Participles, Gerunds, Verbal nouns.
[45] See chapters 2.9.1 Infinitive, 2.9.4 Conditional, 2.9.5 Imperative.
[46] See chapter 3.1.2 Coordination for implementation see 4.11.2 Coordinative composites.
[47] See chapter 2.9.2 Vowels of tense.
[48] In the reality, the rule looks a bit differently:

```
&part /<= &part (¬/)* __
```

This form allows having two (or more) participles in a coordinative composite. Coordinative composite is in fact, two or more separate words connected together – the participle is possible in each of these "subwords". The character / is placed between "subwords" – the automata connected with the rule "forgets" that there was any participle in the previous "subword". See chapter 4.11.1.



The marker &verb in the lexicon verb allows me to write a rule forbidding to form a verb from any stem[49].

The lexicons have following form:

Lexicon verb:

```
\lf |i<&verb>
\lx verb
\alt end
\eng |xInfinitive

\lf |us<&verb>
\lx verb
\alt end
\eng |xKonjunktive

\lf |u<&verb>
\lx verb
\alt end
\eng |xVolitive

\lf |a<&verb>
\lx verb
\alt afterTemp
\eng |xPresent

\lf |i<&verb>
\lx verb
\alt afterTemp
\eng |xPreterite

\lf |o<&verb>
\lx verb
\alt afterTemp
\eng |xFuture
```

The continuation class `afterTemp`:

```
ALTERNATION afterTemp            indicative part
```

Lexicon indicative

```
\lf |s
\lx indicative
\alt end
\eng |xIndicative
```

[49] See chapter 4.6 Category prohibiting rules.

Lexicon part (participles):

```
\lf |nt<&part>
\lx part
\alt afterPart
\eng |xActPart

\lf |t<&part>
\lx part
\alt afterPart
\eng |xPassPart
```

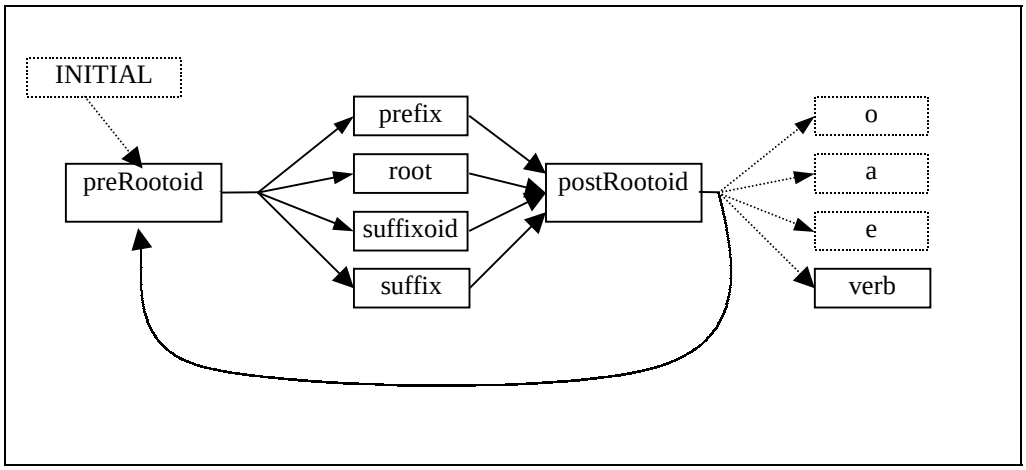
The continuation class `afterPart` was mentioned above.

4.5 Roots

This describes the backbone of the whole system. It covers typical composites, excluding coordinative composites, numbers, etc. The elements are classical roots, most of affixoids and affixes.

As was said in chapter 3.2, affixes are in fact roots. I make only few differences between roots, affixoids and affixes. As was said I make no distinction between prefixoids and prefixes. The main difference between roots on one side and affixes with affixoids on the other is that they are much more used in word building than classical roots. They are also mostly monosyllabic, therefore it is very often possible to analyze a word as a sequence of these small elements, even if it is in fact built from smaller number of longer roots. Other difference is that many of the affixoids are used not fully as a separate root. They very often lack the ability to create all part of speeches.

Because of these reasons, I have created four lexicons – with classical roots[50], with prefixes (prefixoids and true prefixes), with suffixoids and with true suffixes. These lexicons are connected on both sides with organizational lexicons containing only one item each. These items have zero realization. The first lexicon (called `preRootoid`) gives me the opportunity to access all roots in all lexicons, as if they were in one large lexicon. The second (called `postRootoid`) enables me to have one default continuation class for all four lexicons.



The problem of restriction of endings following prefixes is solved by using category prohibiting features[51] for each prefix. They are assigned according to chapter 3.2.3. Prefixes have two types of continuation classes. One class is for the classical prefixes and the other for prefixes that can be used also alone, without any ending (e.g. *fi*, *eks*, *ek*):

```
ALTERNATION afterPrefix                 postRootoid
ALTERNATION afterPrefixAndEnd           postRootoid end
```

In this state of the analyzer, I do not use the inherent categories. However, I can easily imagine that in the next version it would be possible to use them for some restrictions on affixes, for better interpretation of the result or for some module of higher level of linguistic description. For these reasons all roots and affixes have a marker of their category: `¤o`, `¤a`, `¤i` and `¤e` (noun, adjective, verb and adverb). They have all zero surface representation.

[50] The lexicon roots (in file PIV.lex) contains about 11 thousands of roots from the electronic version of the PIV dictionary – see Appendix A.3 Conversion of the PIV.

[51] See chapter 4.6.

4.5.1 Inserted o

As was said in chapter 3.3.1, the letter *o* can be theoretically inserted between any two roots (excluding affixes). In reality, it is inserted only between roots that would be hard to pronounce without it.

I have two possibilities – to allow the inserted *o* between any two roots or to allow it only under some circumstances. I will show rules for both possibilities.

For the first possibility, the only thing I have to ensure is to have roots on both sides of the inserted *o*. Root starts (as any morpheme) with an character `|`, this character is not realized on the surface level. The character `=` is the last character in the root. This character is also realized as a zero (`=:0`) on the surface level. I will allow the realization as an *o* (`=:o`) if it is followed by another root. The only thing that enables to the rule to determine that a sequence of characters is a root, is the `=` at the end of such a sequence. The rule has following form:

```
RULE =:o => __ |:0 (¬|:0)* [=:o | =:0]
```

The expression `|:0 (¬|:0)*` ensures that the character `=` is at the end of the immediately following morpheme.

Another possibility is to allow the *o* only between two consonants. No affix[52] starts (for suffixes) or ends (for prefixes) with a vowel. Therefore is obvious, that if two consonants from different morphemes meet, these morphemes are roots:

```
RULE =:o => C __ |:0 C
```

However, there are also words where the o is for some reasons (tradition, international influence) inserted even after a vowel: (*radio**o**elsendi* – *radiobroadcast*). Such a word contains a character `©` in its features. This character has two possible realizations `0` or `o` (`©:0` or `©:o`).

```
\lf |radi<¤o©>=
\lx root
\alt afterRoot
\eng |ray/radio
```

If the second alternative is chosen, it is good enough to remove the first rule. If the second alternative is chosen, it is also necessary to remove the default realization `©:o`, the default realization `©:0` must be preserved to allow recognizing words as radio, etc. having this character in their lexical entries.

Now, I will show two examples of using two-level rules to restrict some usage of a morpheme. The first example will be prefix *bo* and the second prefix *pra*.

4.5.2 Prefix bo

Prefix *bo* has very restricted usage, it can precede only few selected roots – some family members and few other roots. It would be good to restrict somehow the possibility of assigning the prefix from all roots to these selected only.

One solution would be to create lexicon containing these roots and *bo* would contain link only to this lexicon. Disadvantage of this approach is the fact, that if I would have similar problem with other prefixes, it would require a lexicon for each of them. The problem is that each lexical item can be only in one lexicon, but the required lexicons for prefixes would very likely overlap. This is technically solvable, but the price is high number of small lexicons, complicated continuation classes and need to redesign the system each time some new restricted prefix is added.

The other solution is much easier. I add a marker (`&bo`) to the prefix *bo* and another special symbol (`†bo`) to each root that can accept the prefix *bo*. Then the problem is reduced into the problem of writing a rule, which will allow occurrence of *bo* only if it is followed immediately by the allowed root. To make it easier `&bo` and `†bo` are introduced as single characters. The rule will look like this:

```
RULE &bo => __ 1 (†bo)
```

Symbol `1` is used as an abbreviation for `(¬"|":0)* "|":0 (¬"|":0)*`. The meaning of it is – skip anything in the current morpheme[53], then pass to the next and it is possible to skip anything too, but impossible leave the morpheme. The symbol `1` is used only in this text, in a real rule it has to be inflated into the regular expression it stands for.

---

[52] Except *ĉjo* and *njo*. However, words containing suffixes *ĉjo* and *njo* are handled by separate lexicon entries. Therefore, these suffixes do not participate in word building in my system.

[53] Each morpheme starts with character |. This character is realized as zero on the surface level. To distinguish it from the metacharacter | with meaning "alternative", it is written in quotation marks.

However, this rule does not allow words like *bo|ge|patroj* – *grandparents-in-law*. Therefore, I will allow a morpheme *ge* (both sexes) between the morpheme *bo* and the root with `†bo`:

```
RULE &bo => __ 1 (&ge 1) (†bo)
```

There is another problem with prefix *ge* too. The possible order of the pair of morphemes *bo* and *ge* is fixed. Morpheme *bo* can precede *ge*, but *ge* cannot precede *bo*. The possibility of *bo* before *ge* is incorporated in the preceding rule, the impossibility of the opposite order is ensured by another rule:

```
RULE ge /<= __ 1 &bo
```

4.5.3 Prefix pra

As was said in 3.2.3.4, prefix *pra* has two meanings – with names of relatives, one generation older or younger; with the rest of stems it marks something very old. I treat them as two different prefixes and use rules to prevent undesirable behavior.

The first set of roots is rather small – some of family members. I use the same strategy as with prefix *bo*. The prefix is marked with `&praFam` and possible roots are marked with `†praFam`. The rule looks this way:

```
RULE &praFam => __ 1 †praFam
```

However the prefix can be repeat: *prapraavo* – *great-great-grandfather*. Therefore, I will extend the rule following way (the prefix *pra* has to be immediately followed by root marked with `†praFam` or by another prefix *pra*):

```
RULE &praFam => __ 1 [ †praFam | &praFam ]
```

The roots that can accept the prefix *pra* in the first sense (`&praFam`), cannot accept the prefix in the second sense (marked as `&praPrim`) and the prefixes cannot be combined. Analyses of *praavo* as *primeval grandfather* or of *prapraavo* as *primeval great-grandfather* are impossible. This rule ensures that:

```
RULE &praPri /<= __ 1 [ †praFam | &praFam ]
```

Last thing is to forbid repeating the prefix in the sense primeval:

```
RULE &praPri /<= __ 1 &praPri
```

When the preposition is acting as a normal root (with adjectival or adverbial ending), it has meaning of the something very old; therefore, I have to disable the possibility of assigning these endings to the other *pra*.

Entries for these two prefixes have following form:

```
\lf |pra<&praPri•o•verb>
\lx prefix
\alt afterPrefix
\eng |xPrimeval

\lf |pra<&praFam•o•a•e•verb>
\lx prefix
\alt afterPrefix
\eng |xNextGeneration
```

4.6 Category prohibiting rules

Some stems can have only some category endings, at least in a real text. For example *bo* (see 3.2.3.1) can have only adjective ending – *boa*, and the forms *?boo* or *?boe* are not used. There are two ways how to manage it – by continuation classes or by using rules.

The first possibility is better to use, if it can be applied to some whole set of roots or some type of stems. For example – if all words from lexicon `X` could have adjectival and adverbial endings, but nominal endings were not possible, it would be suitable to create a continuation class `afterX`. This class would contain lexicons `end, a, e` and maybe some other, but not lexicon `o`:

```
ALTERNATION afterX          a e
```

The second possibility is better to use for less compact stems – it would be unsuitable to create thousands of different continuation classes for every different stem. The better opportunity is to use one mark for the ending and another for the stem and then write a rule that will fail if these two markers are together. It is possible to forbid only immediate cooccurences of two elements or even any.

I have created such possibility for the nominal, adjectival, adverbial and verb endings. Each of these endings has a marker (`&o, &a, &e, &verb`) and each of the words that do not want the ending has a marker too (`•o, •a, •e or •verb`). The rule forbids only immediate cooccurences – the following root or affix can totally change the situation. The rule for forbidding noun ending after the root with feature `•o`:

```
RULE &o /<= •o 1 __
```

Symbol `1` is used as an abbreviation for `(¬"|":0)* "|":0 (¬"|":0)*`[54].

The rules for the rest of categories look similarly.

I have created also other two rules – to disable possibility of adding a root (classical roots, without affixes) to the current morpheme. One rule prohibits the immediate adding; one rule prohibits any occurrence of a root. For this purpose, two markers have been introduced: `•root, •neverRoot.` The presence of a classical root can be inferred from the character `=` – the character used for inserting the letter *o* between roots.

```
RULE =:0 /<= •root 1 __
RULE =:0 /<= •neverRoot (¬/)* __
```
[55]

4.7 Personal pronouns

The region of personal pronouns is very easy.

First, it contains lexical entries for all personal pronouns, e.g.:

```
\lf mi
\lx persPronoun
\alt afterPersPron
\eng I
```

Personal pronouns are declined – they can be in nominative or accusative; it is impossible to talk about number (or about other number than singular). Therefore, the continuation class contains link to lexicon case.

Adding the ending *a* to a personal pronoun forms a possessive pronoun. There are two possibilities – to use the adjectival ending *a* or to have a special ending *a*. I used the second possibility. The reason is that a possessive pronoun can be element of a composite: *miaflanke – from/on my side*. Therefore, the ending *a* can be followed by a root. The classical adjectival ending cannot (at least in my model). It could be solved by rules too, but I have chosen this variant.

Therefore, the continuation of a personal pronoun has following form:

```
ALTERNATION afterPersPron        case possesiveA
```

And the lexicon containing possessive ending (the only entry of the lexicon) following one:

```
\lf |a
\lx possesiveA
\alt afterPossesiveA
\eng |possesiveA
```

The possessive ending has two opportunities of realizing itself: to decline, if the pronoun is a separate word, or to be in front of a root, if the pronoun is a part of a composite.

Therefore, the continuation class has this form:

```
ALTERNATION afterPossesiveA      nr preRootoid
```

As was said in chapter 2.6.1, the accusative *sin* is regarded as a separate prefix and is not analyzed as a form of the pronoun *si*.

4.8 Correlatives

I will treat correlatives as simple words – I will not analyze them into their two parts. I will create lexicon containing all 45 forms. Now what about continuation classes.

First, I will deal with declination. The *-iu* (individual) and *-ia* (quality) forms are be fully declined, so their continuation class will contain the lexicon `nr` (lexicon `case` follows lexicon `nr`) and there will not contain the lexicon `end`. The *-io* (thing) and *-ie* (place) can form accusative, so their continuation class will contain the lexicon `case`. The rest of correlatives does not decline, they can be in text without any endings – their continuation classes will contain lexicon `end`.

The traditional forms (*neniaĵo, neniigi, ..*)[56] create a small set that is not going to grow – I will put them as separate lexical entries into the lexicon. These forms cannot participate in further word building, and they even cannot change their part of speech (except participles) – their continuation classes will be direct links to verbal or nominal inflection.

The individual and quality forms can precede many roots. I give up to go through thousands of roots to say which of them are possible and which of them are not. I allow all roots, except true

[54] See chapter 4.5.2.

[55] The character / marks the beginning of a new "subword" in a coordinative composite. Prohibiting of the root is valid only within the "subword". See chapter 4.11.1.

[56] See 2.7.1.3.

suffixes to follow these to types of correlatives – the continuation classes will contain lexicon `preRootoid`.[57]

The form of quantity can be connected with numeral suffixes (as any numeral). Problem is with the possibility to add the *et* or *eg* suffix to diminish or augment the quantity. I cannot simply link the lexicon of suffixes – it contains also other suffixes, which are impossible in this context. To split the suffixes into two lexicons would be quite costly solving for these few words. Another possibility is to link whole lexicon and to disable undesirable suffixes by two-level rules. However, in a real text occur only forms derived from *iom – some quantity*: *iomete* and *iomege*. Therefore, the best solution is to insert these two forms directly into the lexicon.

The rest is easy. The forms of place, time and manner can form adjectives. The forms of quantity can form adverbs.

The lexicon correlative has following form (the cont. classes are defined at the end)

```
\lf tia
\lx correlative
\alt afterCorrelIa
\eng such

\lf tial
\lx correlative
\alt end
\eng so

\lf tiam
\lx correlative
\alt endA
\eng then

\lf tie
\lx correlative
\alt afterCorrelIe
\eng there

\lf tiel
\lx correlative
\alt endA
\eng thus

\lf ties
\lx correlative
\alt end
\eng thatOnes

\lf tio
\lx correlative
\alt case
\eng that

\lf tiom
\lx correlative
\alt afterCorrelIom
\eng thatMuch

\lf tiu
\lx correlative
\alt afterCorrelIu
\eng thatOne
```

[57] `Prerootoid` is an organizational lexicon containing only one item with zero realization. This lexicon enables me to access all roots in many different lexicons, as if they were in one large lexicon. See chapter 4.5 Roots.

The part containing nonanalyzed forms:

```
\lf kial
\lx correlative
\alt o
\eng reason

\lf tieul
\lx correlative
\alt end
\eng [|tie|ul]man from there

\lf iele
\lx correlative
\alt end
\eng [|iel|e]emphasized somehow
```

kiele looks the same

```
\lf iomet
\lx correlative
\alt oAE
\eng [|iom|et - |someQuantity|xDiminish]a bit
```

iomeg(e) looks the same

The part containing traditional forms:

```
\lf neniig
\lx correlative
\alt verb
\eng [|neni|ig - |nothing|xToCauseOrLetToDo]destroy
```

neniiĝ(i) looks the same

```
\lf neniajx
\lx correlative
\alt o
\eng [|neni|ajx - |nothing|xThing]nearlyNothing

\lf neniec
\lx correlative
\alt o
\eng [|neni|ec - |nothing|xAbstractQuality]nothingness

\lf tiajx
\lx correlative
\alt o
\eng [|tia|ajxc - |such|xThing]suchThing
```

The continuation classes (classes with the same name as lexicon they link to, are not listed):

```
ALTERNATION afterCorrelIu               nr preRootoid
ALTERNATION afterCorrelIa               nr preRootoid
ALTERNATION afterCorrelIe               case a
ALTERNATION afterCorrelIom              end numSuffix e
ALTERNATION endA                        end a
```

4.9 Numbers

I have a possibility to describe numbers thoroughly, according to the expression in chapter 2.8.1. However, complicated numbers expressed by words are very rare in a real text (none in my corpus). Therefore, it would be too much work for a small result. I handle many aspects of a number, but I have also left some uncovered.

1) I will start with simple numerals. Except the numeral *unu* – *one*, they cannot be declined. I have allowed also the unofficial form *unun*. Therefore, the declension of the numeral *unu* is ensured simply by having the lexicon `nr` in its alternation class. The other numerals have to have the lexicon `end` in their continuation classes.

2) Numerals 2 to 9 can be joined with *dek* – *10* or *cent* – *100*, to make a multiple. Therefore, numerals 2 to 9 have in their continuation classes link to the lexicon with numerals 10, and 100 – `numDekCent`. A compound cardinal numeral is a sequence of simple numerals or these multiples. Each word is parsed separately.

3) Numerals can be followed by a numeral suffix (*obl*, *op*, etc.) or a category ending. If it is a simple numeral, the suffix (or ending) is simply added to it. However, if it is a compound numeral, the spaces have to be replaced by hyphens. This hyphen is not possible for cardinal numerals.

This fact can be easily handled by a rule. The rule will allow presence of a hyphen only if the numeral is followed by a suffix or ending. I have two lexicons with one entry each. The first lexicon has an entry for the hyphen. The entry of the second lexicon precedes all possible morphemes that can be added to the numeral. Each entry has a marker. The one of the first lexicon is allowed only, if it is followed (not immediately) by the marker of the second lexicon.

The lexicon with the hyphen:

```
\lf &NumHyphen-
\lx numHyphen
\alt afterNumHyphen
\eng _
```

The lexicon, that is before a suffix or an ending:

```
\lf &NumHyphOk
\lx numHyphOk
\alt afterNumHyphOk
\eng _
```

The rule that allows a hyphen only together with a suffix or an ending:

```
RULE &NumHyphen => ___ (¬/:0)* &NumHyphOk
```

The continuation class `afterNumHyphen` contains links to lexicons with numbers 1 to 9, 10, 100, 1000:

```
ALTERNATION afterNumHyphen        num1To9 numDekCentMil
```

The continuation class `afterNumHyphOk` contains links to suffixes (`numSuffix`), to endings (`o` and `e`).

```
ALTERNATION afterNumHyphOk
                                  o e numSuffix preRootoid
```

The numeral suffix can be followed by anything that the classical suffix can (category ending, root, affix). I will use the same continuation class.

```
ALTERNATION afterSuffix           postRootoid
```

For example the suffix for creating fractions:

```
\lf |on
\lx numSuffix
\alt afterSuffix
\eng |xFraction
```

4) Ordinal numbers are expressed by adding the ending *a*. This ending can be considered to be an adjectival ending. However, it has a bit different properties – it is not removed in composites. Therefore, I have created a lexical entry containing this ending (for practical reasons, the ending is put into the lexicon `numSuffix`):

```
\lf |a
\lx numSuffix
\alt afterOrdinal
\eng |xOrdinal
```

An ordinal number has to be declined; therefore, the continuation class `afterOrdinal` contains link to the lexicon `nr`. It can be also part of a composite – it can be followed by a root or an affix. However, it is impossible to add a category ending to it. Therefore, the continuation class contains also lexicon `preRootoid`:

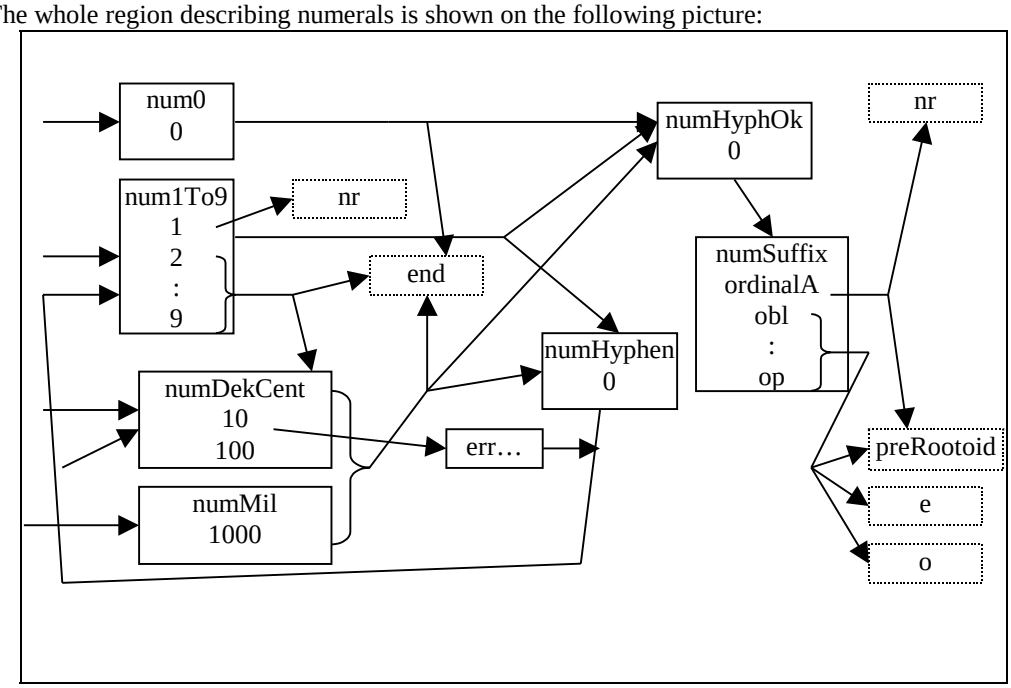
```
ALTERNATION afterOrdinal                nr preRootoid
```

5) As was said, between exponents of a cardinal compound numeral, there have to be spaces. On the same place of other numerals is a hyphen. However in a real text this rule is very often violated and the numbers a rewritten as one word. This case is very often (82 times in my corpus) with numbers lower then hundred (**kvindekkvarfoje* – *kvindek-kvarfoje* – *fifty four times*), and especially with

numbers between ten and twenty (**dekdu*, *instead of dek du – 12*). I have allowed parsing these numerals with a remark, that an error occurred. Therefore, it is added lexicon `errNumTogether` to the continuation class of the lexical entry 10. This lexicon contains only one item. The item has a zero surface realization, and a gloss with the description of the error:

```
\lf 0
\lx errNumTogether
\alt afterErrNumTogether
\eng |##Error - must be writen separately##
```

The continuation class of this entry points to the lexicon `num1To9`.
The whole region describing numerals is shown on the following picture:



All continuation classes have following form:

```
ALTERNATION num                    num0 num1To9 numXCM
ALTERNATION afterNum0              end numHyphOk
ALTERNATION afterNum1              nr numHyphen numHyphOk
ALTERNATION afterNum2To9
                    end numXCM numHyphen numHyphOk
ALTERNATION afterNumDek
                    end numHyphen numHyphOk errNumTogether
ALTERNATION afterNumCent           end numHyphen numHyphOk
ALTERNATION afterNumMil            end numHyphen numHyphOk
ALTERNATION afterNumHyphen         num1To9 numDekCentMil
ALTERNATION afterNumHyphOk         o e numSuffix preRootoid
ALTERNATION afterOrdinal           nr preRootoid
ALTERNATION afterErrNumTogether    num1To9
```

4.10 Countries

I have created two lexicons `country` and `inhabitant`. The words in the lexicon `country` are names of countries, and the names of inhabitants are created by the suffix *ano* (*Peruo* → *Peru|ano* – *Peru*). The words in the lexicon `inhabitant` are names of the nationalities and countries are derived by suffixes *io*, *ujo*, *lando* and *istano* (*Italo* → *Ital|io*). The names derived from the name of a town or a river by suffix *io* (*Alĝero* (*Algiers*) → *Alĝerio* (*Algeria*)) are considered to be primitive and have their entries in the lexicon `Country`.

Country

The only problem with the first group are the names in plural (*Bahamo**j*** – *Bahamas*). However, I have allowed plural also for names, therefore these countries are also recognized. It would be possible to use two entries for each such country – one unanalyzed:

```
\lf bahamoj
\lx country
\alt case
\eng [|baham<¤o>|xNoun|xPlural]Bahamas
```

And one with the root with prohibited nominal ending (`•o`):

```
\lf bahamoj<¤o•o>
\lx country
\alt afterCountry
\eng Bahamas
```

However, I have not done that. I left it for the next version. The plural with names should be treated systematically. It is possible to write a rule that prohibits plural ending with some roots. It should be prohibited also after *lando* (*land*) in names of countries like *Pol|lando* – *Poland*, however the word *lando* in other occurrences can form the plural.

The typical entry of this lexicon:

```
\lf argentin<¤o>
\lx country
\alt afterCountry
\eng Argentina
```

With continuation class:

```
ALTERNATION afterCountry         postRootoid
```

Inhabitant

The roots in the lexicon `inhabitant` are separated into three parts:

1) Forming the name of the country by the suffixes *io* or *ujo*.
   Hungaro → *Hungario*/*Hungarujo*, *Turko* → *Turkio*/*Turkujo*
2) Forming the name of the country by the adding the root *lando*.
   Finno → *Finnlando*, *Skoto* → *Skotlando*
3) Forming the name of the country by the suffix *istano*.[58]
   Uzbeko → *Uzbekistano*, *Afgano* → *Afganistano*

The countries in the subgroup 2 and 3 are easy to handle – *land* is a normal root in the main lexicon[59], *istan* is in separate lexicon. The continuation classes have following form:

```
ALTERNATION afterInhabitantLand         postRootoid
ALTERNATION afterInhabitantIstan        postRootoid
istan
```

There are two problems with the first subgroup:

1) The suffix *ujo* is used with other words as a suffix with meaning of a container.
2) The suffix *io* is used with some words as a suffix with meaning of a science.
3) Word *patro* – *father*, can take suffixes *ujo* and *io* to form *patrio*/*patrujo* – *"one's own country"*.

I have created four different suffixes and three markers: *ujo* (`&iUjCtr`) for a country, *ujo* (`&iUjBox`) for a container, *io* (`&iUjCtr`) for a country and *io* (`&iSci`) for a science. All roots that can form a country by the suffixes *ujo* or *io* have another marker (`†iUjCtr`). The possible cooccurences are treated by rules:

1) Suffixes *io* or *ujo* can be applied only if the previous root has a marker `†iUjCtr`
   `RULE &iUjCtr => †iUjCtr 1 __`
2) An inhabitant cannot be put into a box:
   `RULE &iUjBox /<= †iUjCtr 1 __`
3) Suffix *io* for creating sciences is allowed only with special roots (see 4.11.1).

The typical entries of this lexicon:

```
\lf sloven<¤o†iUjCtr>
\lx inhabitant
\alt afterInhabitantIUj
\eng Slovene
```

[58] *Pakistano* belongs to the lexicon country, the inhabitant is called *Pakistan|ano*.

[59] Maybe it would be reasonable to use new suffix *lando* (it has a bit different meaning that the word *lando*).

```
\lf skot<¤o>
\lx inhabitant
\alt afterInhabitantLand
\eng Scotsman

\lf afgan<¤o>
\lx inhabitant
\alt afterInhabitantIstan
\eng Afghan
```

The entries have different continuation classes:

```
ALTERNATION afterInhabitantIUj   postRootoid
ALTERNATION afterInhabitantLand  postRootoid
ALTERNATION afterInhabitantIstan postRootoid istan
```

4.11 The rest

4.11.1 Sciences

As was said in the chapter 3.3.3, the structure of names of sciences is rather complicated.

I think that it is reasonable to use the suffix *io*. I have used the same strategy as with the same suffix creating names of countries from their inhabitants[60]. The lexical entry of these two suffixes are, of course, separate. I have introduced a marker `&iSci` for marking the suffix and another marker `†iSci` for marking the roots, which can accept the suffix. The rule has following form:

```
RULE &iSci => †iSci 1 __
```

The marker has received also suffix *ologo:*

```
\lf |olog<¤o†iSci>
\lx suffixoid
\alt afterSuffixoid
\eng |xScientist
```

The typical entry of a scientist, that can take suffix *io*:

```
\lf |paleontolog<¤o†iSci>=
\lx root
\alt afterRoot
\eng |paleontologist
```

The rest is managed by separate lexical entries – I do not consider the *iko* to be a morpheme, scientists with the pseudosuffix *olog* have also their separate entry. Words that do not form the pair scientists – science (*ideologo* – *ideologist*, *ideologio* – *ideology*) have their own entries and are not regarded as derived one from the other.

4.11.2 Coordinative composites

Coordinative composites are treated very simply – I have allowed to any word to be followed by any other word separated by a hyphen.

The coordinative composites without a hyphen and without an ending in the middle of them (*nordoriento*) are recognized by the part for normal determinative composites. The composites without a hyphen and with an ending (*nigrablanka*) are not recognized.

When the word is complete (all grammatical endings are assigned or are not necessary) it is followed by the lexicon `end.` This lexicon has two entries – first allows finishing the word, the second allows to insert a hyphen and continue with the second word of the coordinative composite:

```
\lf 0
\lx end
\alt #

\lf -/
\lx end
\alt start
\eng –
```

[60] See chapter 4.10.

Character / enables to the rules to distinguish the end of the "subword", reset ("forget") and start checking the second "subword". For example, there can be only one participle suffix in a word[61], however it is possible to put two participles into coordination. The rules than has following form:

```
&part /<= &part (¬/)* __
```

4.11.3 Prepositions

Prepositions are very productive. They can stay alone, it is possible to attach any category ending to them and they can act as prefixes. I use a general continuation class `afterPreposition` allowing all these possibilities. Exceptions are handled by rules for category prohibiting[62].

The typical entry has following form:

```
\lf |anstataux
\lx preposition
\alt afterPreposition
\eng |insteadOf
```

The main work is done by the continuation class. It is obvious that the continuation class has to content the lexicon `end` (the preposition as it is). Most of the prepositions can have all category endings, therefore there must be lexicons `o, a, e` and `verb.` There must be also lexicon `preRootoid` to allow to the preposition to act as a prefix. Last lexicon is `prepositionableWords` – some words can be also prefixed by a preposition (*nun* – *now* → *ĝisnuna* – *lasting till now*). The whole continuation class:

```
ALTERNATION afterPreposition
        end preRootoid prepositionableWord a e o verb
```

In the next version of the analyzer, it would be good to apply some restriction onto the prefixation – it is for example possible to create words mixing adverbs of place with preposition of time (**apudnuna* – *beside-now* (adjective)).

Few prepositions do not accept all category endings (at least normally). For example, *al* – *to* does not form a noun (**alo*)[63]. This can be ensured by category forbidding rules[64]. The only thing that is necessary to do on this place, is to add a specific marker (in this case `•o`) to the features of the preposition:

```
\lf |al<•o>
\lx preposition
\alt afterPreposition
\eng |to
```

4.11.4 Primitive words

Primitive words (words that do not need a grammatical ending), are put as separate entries into the specific lexicons. Numbers and correlatives are described in different chapters[65].

To this category belong conjunctions, interjections, some adverbs and particles.

4.11.5 Suffixes ĉj/nj

Words with suffixes *ĉj* and *nj* are treated as separate lexicon entries. These words cannot be part of any composition (except coordinative compositions).

The typical entry of this lexicon has following form:

```
\lf panj
\lx cxjonjo
\alt oAE
\eng [|patr|in|nj]mum
```

4.11.6 Units

Units and scientific prefixes are separated from the rest and cannot be mixed with other roots. There are two lexicons – units and SI prefixes. The unit with no prefix is considered to be with zero prefix.

---

[61] See chapter 4.4.
[62] See chapter 4.6.
[63] There is a word *alo* – wing, but it has nothing to do with the preposition *al.*
[64] See chapter 4.6.
[65] For correlatives see 4.8, for numbers see 4.9.

The word *metro* in meaning *measuring device* is considered to be a different root and is in the main lexicon as any other root. Word *ampermetro* – *ammeter*, *voltmetro* – *voltmeter* and *om(o)metro* – *ohmmeter* are in the lexicon as separate entries:

```
\lf |om©metr<¤o>=[66]
\lx root
\alt afterRoot
\eng [|om<¤o>=|metr<¤o>=]ohmmeter
```

The impossibility to mix units with other roots has another consequence – the unit can be preceded by the root *kub(o)* – *cube* and *kvadrat(o)* – *square*: *kubmetro* – *cubic meter*. This problem is provisionally solved by putting these two words into the lexicon containing `SIPrefix` with linkage back to `SIPrefix` lexicon (containing zero entry)

4.11.7 Replacing ĥ after r by k

As was said in chapter 2.1, any *ĥ* after *r* in the same morpheme can be replaced by *k*. It is very easy to write such a rule:

```
RULE hx:k => r __
```

[66] For explication of the character ©, see chapter 4.5.1 Inserted o.

5 Conclusion

The system was tested on set of Esperanto texts containing about 460 000 words and has covered about 97.5 % of them.

Most of the unanalyzed words are proper names or misspellings. I have not inserted names specific for the text – as family names, names of small cities, etc. If I inserted 10 most common names (Rikita, Kefalín, Saturnin, Vaněk, Vlaĥ, Vilík, Barunka, Viktorka, Roy, Kristla) with 2308 occurrences into the lexicon, the analyzer would recognize 98 % of the text. The number looks very nice, however it is partly implication of the corpus structure[67]. For a real corpus (newspapers, spoken text, original texts written by people from different nations, etc.) the number would be not so good. However, in my opinion the decrease of coverage would be caused mostly by the large amount of different proper names, and not by common words.

Such a high number with a small lexicon of about 11 thousands entries is a consequence of Esperanto rich word building. Many words that could not be regarded as derived in other languages (at least from synchronic point of view) and would require separate lexical entries (*town – mayor – city*, *father – mother*, *house – hovel*), are derived in Esperanto from one root (*urbo – urbestro – urbego*, *patro – patrino, domo – domaĉo*). Therefore, the size of the lexicon can be substantially smaller.

The disadvantage is that many words can have more than one analysis (I am not talking about grammatical homonymy). There is a large set of affixes – very often used short morphemes. Moreover, as was shown in the chapter 3, nearly all cooccurrences of various morphemes are allowed. The only limit is the fantasy of the Esperanto speaker. I list few examples:

```
dirite
di<¤o©>=|rit<¤o>=|e<&e>              |god|rite|xAdverb
dir<¤i>=|i<&verb>|t<&part>|e<&e>    |say|xPreter|xPassPart|xAdverb

doktoro
dok<¤o>=|tor<¤o>=|o<&o>      |dock|torus|xNoun (?!)
dokt<¤a>=|or<¤o>=|o<&o>      |erudite|gold|xNoun (?!)
doktor<¤o>=|o<&o>            |doctor|xNoun

avineto
av<¤o†bo†praFam>=|in<¤o†bo>|et|o<&o>
                        |grandfather|xFeminine|xDiminish|xNoun
avi<¤o>=|net<¤a>=|o<&o>      |airplane|precise|xNoun

papero
pap<¤o>=|er<¤o>|o<&o>      |pope|xElement|xNoun – element of a pope (?!)
paper<¤o>=|o<&o>           |paper|xNoun

ili
il<¤o>|i<&verb>      |xTool|xInfinitive – to tool
li     they
```

Ridiculous analysis of *papero* as *an element of a pope* could be prevented by prohibiting assigning the affix *er* to countable nouns. This approach was used for some simple things (prefix *pra* or *bo*) and could be used more generally. However, the classification of roots is very time consuming if the feature has to be assigned to a large set of roots. It would also require further study and analysis of a large amount of texts. Two level rules were used mostly together with such a classification for this – there are no phonological alternations in Esperanto.

On the contrary, inflection is totally unambiguous. Therefore the total number of ambiguities is not so high (13.64 %).

There are still some areas to cover – especially proper names, their capitalization and connection to Esperanto inflection. It would be good to allow recognition of common mistakes. This was implemented for numbers, however it could be used for unofficial names of countries, some composites with correlatives, some common misspellings (use of u instead of ŭ), common errors resulting from scanning (e.g. *m* → *rn*), etc. Other question is adapting the system for using as a reasonable generator.

[67] See chapter Resources – Corpus

The used program could be also improved:

1) Incremental recognition (first most common things, than rest, than mistakes).

2) Possibility to add probabilities to continuation classes and rules. The result of the analysis would be sorted according the product of these probabilities.

3) Possibility to use list of lexicons instead of a continuation class in a lexicon entry.

4) Better connection between composites that are whole in the lexicon and their parts.

5) Better integration with other tools. This could be used for connection with a unification grammar (the current grammar is very simple).

6) Unicode support.

Resources

The two most commonly used sources are referred through the text by the abbreviations PAG for Plena Analiza Gramatiko and PMEG for Plena Manlibro de Esperanto. The PAG is followed by the paragraph number, PMEG by the name of the html page. Sources of examples are marked with a little superscript at the end of the example – A for PAG, M for PMEG and H for examples from the grammar overview of the dictionary by Rudolf Hromada. The examples in these grammars are very often taken from some real texts, mostly from texts written by Zamenhof.

If necessary, the original title is followed by English translation in Italics.

PAG – Kálmán Kalocsay, Gaston Waringhien: Plena Analiza Gramatiko de Esperanto, *The full Analytical Grammar of Esperanto*, Universala Esperanto-Asocio, Rotterdam 1985

PMEG – Bertilo Wennergren: PMEG, Plena Manlibro de Esperanta Gramatiko, Versio 8, *The full manual of the Esperanto Grammar, Version 8*, 1998, http://purl.oclc.org/NET/pmeg

Frequently Asked Questions (FAQ) for soc.culture.esperanto and esperanto-l@netcom.com from 1998-04-21

Antworth Evan L.: User's Guide to PC-KIMMO Version 2, Summer Institute of Linguistics, 1995. *http://www.sil.org/pckimmo/v2/doc/guide.html*

Barandovská Věra: Esperanto pro samouky, *Esperanto teach-yourself*, SPN, Praha 1989

Filip Jan, Filip Karel: Velký slovník česko-esperantský, *The Grand dictionnary Czech-Esperanto*, Slovenský esperanský svaz – INKLEC, Praha 1989 (reprint from 1949)

Harlow Don: Word-Building with Esperanto Affixes, 1995, http://www.webcom.com/~donh/Esperanto/affixes.html

Harlow Don: The Esperanto Correlatives, http://www.webcom.com/~donh/Esperanto/correlatives.html

Hromada Rudolf: Esperantsko-český a česko-esperantský kapesní slovních, *The Pocket-book Dictionary Esperanto-Czech and Czech-Esperanto*, Český esperantský svaz, Praha 1989

Koskenniemi Kimmo: Two-level morphology: A general computational model for word-form recognition and generation, Publication No. 11. Helsinki: University of Helsinki, Department of General Linguistics 1983

Kraft Karel: Česko-esperantský slovník/Ĉeĥa-esperanta vortaro, *The dictionary Czech-Esperanto*, KAVA-PECH Dobřichovice 1998

Kraft Karel, Malovec Miroslav: Esperantsko-český slovník/Esperanta-ĉeĥa vortaro, *The dictionary Esperanto-Czech*, KAVA-PECH Dobřichovice 1995

J.M.D. Meiklejohn: The English Language – Its grammar, history and literature, London 1895

Neal McBurnett: list of English words with Esperanto translation, gopher://wiretap.spies.com/0Library/Article/Language/esperant.eng

Microsof Bookshelf 1994, Microsoft Corporation. at CD-ROM, I have used these parts:
Funk and Wagnall's The World Almanac
The American Heritage Dictionary of the English Language, Houghton Mifflin Company, 1992.
Roget's Thesaurus of English words and phrases Longman Group UK Ltd. 1987.
The Concise Columbia Encyclopedia, Columbia University Press 1991

Oficialaj Informoj de la Akademio de Esperanto, n-ro 9, La Letero de l' Akademio, n-ro 7, Aprilo - Majo - Junio 1989.

Petr Jan et al.: Mluvnice češtiny, *The Grammar of Czech Language*, Academia, Praha 1986

Plena Ilustrita Vortaro (PIV) in electronic version (only entry headings), adapted by Klaus Schubert from BSO/Research, ftp://ftp.stack.nl/pub/esperanto/word-lists.dir/piv.tar.Z

Terry L. Smith: The Building Blocks of Esperanto, http://osprey.unf.edu/faculty/tsmith/esp/index.html

The program PC-Kimmo is freely available by Summer Institute of Linguistics at *http://www.sil.org/pckimmo/v2*

Corpus:

The "corpus" for testing the system consists of about 460 000 words. It is not an ideal corpus – there are no newspapers, most of the books are translations, most of them from Czech and five by the same person. However, for as a first version of the morphological analyzer, it fulfilled its purpose.

For conversion of texts to the format acceptable by the PC-Kimmo, I developed a small program described in Appendix A.1. Some of the texts contained a small dictionary at the end – the dictionaries were not included into the corpus.

Except the texts by H.C. Andersen, speech of L. Zamenhof and the novel by U. Matthias, all the texts were donated by Petr Chrdle, the owner of a publishing house KAVA-PECH, Dobřichovice, Czechia, to whom I am really grateful.

List of the texts in the corpus:

H.C. Andersen: Post jarmiloj (translation by L. L. ZAMENHOF)
H.C. Andersen: Anneto (translation by L. L. ZAMENHOF)
H.C. Andersen: Infana babilado (translation by L. L. ZAMENHOF)
H.C. Andersen: Peco da perlovico (translation by L. L. ZAMENHOF)
H.C. Andersen: Plumo kaj inkujo (translation by L. L. ZAMENHOF)
H.C. Andersen: Pupludisto (translation by L. L. ZAMENHOF)
H.C. Andersen: Du fratoj (translation by L. L. ZAMENHOF)
H.C. Andersen: Malnova preĝeja sonorilo (translation by L. L. ZAMENHOF)
H.C. Andersen: Dekdu per la poŝto (translation by L. L. ZAMENHOF)
H.C. Andersen: Sterkskarabo (translation by L. L. ZAMENHOF)
H.C. Andersen: Kion la patro faras, estas ĉiam ĝusta (translation by L. L. ZAMENHOF)
H.C. Andersen: Neĝulo (translation by L. L. ZAMENHOF)
All at: http://www.best.com/~donh/Esperanto/Literaturo/
Karel Čapek: Libro de apokrifoj (translation by Josef Vondroušek)
Václav Chaloupecký: Karolo la IV-a kaj Bohemio (translation by Josef Vondroušek)
Anton Pavlovich Chehkov: Urso (translation by Josef Vondroušek)
Jaroslav Foglar: La Knaboj De La Kastora Rivero (translation by Dieter Berndt)
Václav Havel: Audienco (translation by Josef Vondroušek)
Zdeněk Jirotka: Saturnin (translation by Josef Vondroušek)
J.A. Komenský: Labirinto de la mondo kaj paradizo de la koro (translation)
Ulrich Matthias: Fajron sentas mi interne, ftp://ftp.stack.nl/pub/esperanto/incoming/Fajron.txt
Božena Němcová: Avineto (translation by V. Tobek and K. Procházka)
George Bernard Shaw: Homo de la destino (translation)
Vladimír Škutina: La malliberulo de prezidento (translation by Marie Bartovská).
Miroslav Švandrlík: La nigraj baronoj (translation)
Bruno Vogelmann: La Nova Realismo (first two parts)
The speech of L.L. Zamenhof at 3rd Esperanto Congress in Cambridge, 12th august 1907, at ftp://ftp.stack.nl/pub/esperanto/esperanto-texts.dir/parol.zip
Three articles from La Nica Literatura Revuo n-ro 14, 1958, p. 115-120; n-ro 16, 1958, p. 147-148,; n-ro 27, 1960, p. 115-120

Appendix A Auxiliary programs

During the development of the system, I have used many auxiliary programs. I add remarks only for main three programs. All programs were written in Java using Sun JDK 1.1.6. The resulting program is rather slow when compared to the same type of program written in C++. However, the development and maintenance time is significantly lower.

All programs have very low level of robustness, they are very often unprepared for undesired input. The user interface is also very rough, everything must be done from the command line, no menus and no windows.

Appendix A.1 Conversion to corpus

For the tests of the system, I have used various Esperanto texts. It was necessary to prepare change these texts into the format suitable for the program PC-Kimmo and my system.

Original texts used different type of encoding of characters with diacritics (Latin3 or Latin2, eventually special characters after the letter with the diacritics, e.g. s^ for ŝ, e~ for ě, etc.). It would be good to add recognition of other encodings (various Unicode encoding, Kamenicky, etc.)

My program uses the x[68] encoding of Esperanto texts. I do not use very much of other accented characters – except some very often used. Most of the texts where written by Czechs, so it uses often Czech characters. If these characters are part of the Western character set (á, é, í, ó, ú, ý and š), I use them as separate characters. The rest is replaced by a pair of character without diacritics followed by some special symbol (~ for hacek, ° for circle). It would be possible to add a large amount of other accented character, however it would decrease the speed of the analysis, therefore I have not done it.

Usage of the batch file toCorpus.bat

To make handling with the Java program easier, a batch file is used. The batch file requires as a parameter so called list file. The name of the list file has to be in form *ToCorpus.list, however only the part represented by asterisk is passed to the batch file as parameter

Command

```
toCorpus av
```

will convet the files as stated in the file avToCorpus.list.

Format of the list file

List file specifies names of files to be converted, used encoding and input and output folder. The format of the list file is following:

- `<folder` – input file for all files listed between this line and another line of the same format is set to the specified folder.
- `>folder` – output file for all files listed between this line and another line of the same format is set to the specified folder.
- `:encoding` – the encoding for all files listed between this line and another line of the same format is set to specified encoding. Possible values are `Latin3` and `Latin2^`.

  `Latin3` is used for files using Latin 3 encoding (only Esperanto letters are converted).

  `Latin2^` (default) is used for files using character ^ or x after the Esperanto accented letter and Latin2 codes for Czech letters. The character ~ for hacek or circle and ° for circle is also possible.
- # switches an html encoding on and off. Default is off. When the html encoding is switched on, all html tags (things in < > brackets) are left out.
- `file` – name of the file to be processed. The name is stated without extension. The extension of the input file has to be ".html" if the html encoding is switched on, and ".txt" otherwise.

Example of a list file:

```
<D:\diplomka\Corpus\X\orig
>D:\diplomka\Corpus\X\in
:Latin3
```

[68] See chapter 2.1 Writing and pronunciation.

```
Prope_
:Latin2^
apokrifo
<D:\diplomka\Corpus\avineto\orig
>D:\diplomka\Corpus\avineto\in
avineto
```

Appendix A.2 Filtering result of analysis

The result of the PC-Kimmo command `file recognize` is a file with a sequence of line separated groups each consisting of a surface form by one or more lexical forms. If the analysis of the surface string failed it is followed by line with text "*** NONE ***". The alignment must be off (command `set alignment off`).

I have created a program that can parse this output file and creates three files – None.txt, More.txt and Summary.txt. The None.txt file contains all surface string that the analysis failed on, the More.txt file contains all surface strings with more than one analysis (the analyses follow their surface form). Entries of these files are sorted by the frequency in the corpus, and entries with the same frequency are sorted by alphabet. The summary.txt file contains information about totally parsed words, number of words with no analysis and number of words with more than one analysis, these number are followed by numbers of distinct words and by percentage.

These output files are useful for tuning the system. However, it is necessary to have in mind that it is possible to have an word with a analysis, however the analysis can be bad.

Usage of the batch file filterBad.bat

To make handling with the Java program easier, a batch file is used. The batch file requires as a parameter so called list file. The name of the list file has to be in form `*filterBad.list`, however only the part represented by asterisk is passed to the batch file as parameter

Command

```
filterBad av
```

will parse the files as stated in the file `avFilterBad.list`.

Format of the list file

List file specifies names of files to be parsed, and input and output folders. The format of the list file is following:

- `<`*`folder`* – input folder for all files listed between this line and another line of the same format is set to specified folder.
- `>`*`folder`* – output folder for all files listed between this line and another line of the same format is set to specified folder.
- `file` – name of the file to be parsed. The name includes the extension.

Example of a list file:

```
>D:\diplomka\Corpus\bad
<D:\diplomka\Corpus\x\out
apokrifo.txt
<D:\diplomka\Corpus\avineto\out
avineto.txt
<D:\diplomka\Corpus\saturnin\out
saturnin.txt
```

Appendix A.3 Conversion of the PIV

As a source for the main lexicon of roots, I have used the electronic version of the Plena Ilustrita Vortaro de Esperanto (PIV). I have converted it to the format suitable for PC-Kimmo and merged with the English-Esperanto dictionary based on the dictionary written by Neal McBurnett. Words that are not in the English-Esperanto dictionary have a question mark as the English gloss.

This conversion and merging was done automatically. However, a large amount of changes was done by hand. PIV makes no distinction between ŭ and u, therefore I went through all words

contain *au*, *ou* or *eu* and corrected them[69]. Many lexicons were created totally by hand with PIV and PAG as a source – units, *njo*/*ĉjo* words, primitive words, numbers, affixes, interjections and all names.

Usage of the java class parsePIV.class

The Java application class requires as a parameter a file with information about the name of the PIV files, names of the English-Esperanto dictionary and output folder. The command can have following form:

```
java -classpath %CLASSPATH% parsePIV parsePIV.list
```

Format of the list file

List file specifies names of files to be parsed, and input and output folders. The format of the list file is following:

- `<folder` – input folder for all files listed between this line and another line of the same format is set to specified folder.
- `>folder` – output folder for all files listed between this line and another line of the same format is set to specified folder.
- `:dicFile` – a full name of the file containing the English-Esperanto dictionary. Each line of dictionary file contains an English word, followed by a tab character, followed by Esperanto word.
- `file` – names of the files of the PIV dictionary to be parsed. These files must contain one Esperanto word on each line[70].

Example of a list file:

```
<D:\Diplomka\piv
>D:\Diplomka\x
:D:\Diplomka\X\EngEsr.txt
a.min
b.min
c.min
:
v.min
z.min
```

[69] With help of the Kraft: Esperantsko-český slovník. Roots that were not found in this dictionnary, use *ŭ* (because it is very more often in middle of a root) and have an line `\ux ?` to be easily found.

[70] The electronic version of PIV was adapted by Klaus Schubert from BSO/Research to the following format:

1) Original files with the lexical entry followed by usage remarks and examples of derived words. All letters are capitals. Files do not have extension.

2) Files containing list of common names, proper capitalization (*.min files).

3) Files containing list of proper names, proper capitalization (*.maj files).

I have used only the set with common names, proper names were entered manually.

Appendix B Output of the analysis

Appendix B.1 Sample of the morphological analysis

Source text

Longe, longe ĝi jam estas, kiam mi la lastan fojon rigardis en tiun amindan kvietan vizaĝon, kiam mi kovris per kisoj tiujn palajn, sulkoplenajn vangojn, enrigardadis la bluan okulon, en kiu vidiĝis tiom da boneco kaj amo; longe ĝi estas, kiam min je lasta fojo benis ŝiaj maljunaj manoj!

It is long, long time ago when I looked for the last time in that lovely quite face, when I covered with kisses these pale creasy cheeks, looked into blue eyes, in which could be seen so much godness and love; it is long time, when I for the last time blessed her old hands!

Se mi scius majstre per peniko labori, mi vin glorus, kara avineto, alie; sed mi ne scias, ne scias, kiel tiu ĉi skizo plume desegnita al iuj ekplaĉos!

If I knew to work with a brush as a master, I would commemorate you, in other manner, however I don't know, I don't know how this sketch painted by pen will seam nice to anybody!

Result of the analysis

```
longe
|long<¤a>=|e<&e>     |long|xAdverb

longe
|long<¤a>=|e<&e>     |long|xAdverb

gxi
gxi     it

jam
|jam     already

estas
|est<¤i>=|a<&verb>|s     |be|xPresent|xIndicative

kiam
kiam     when

mi
mi     I

la
|la     the

lastan
|last<¤a>=|a<&a>|n     |last|xAdjective|xAccusative

fojon
|foj<¤o>=|o<&o>|n     |occasion|xNoun|xAccusative

rigardis
|rig<¤i>=|ard<¤i>=|i<&verb>|s     |rig|glow|xPreterite|xIndicative
|rigard<¤i>=|i<&verb>|s     |look|xPreterite|xIndicative

en
|en     |in

tiun
tiu|n     thatOne|xAccusative

amindan
|am<¤i>=|ind<¤a>|a<&a>|n     |love|xDeserving|xAdjective|xAccusative
```

```
kvietan
|kviet<¤a>=|a<&a>|n     |quiet|xAdjective|xAccusative

vizagxon
|viz<¤o>=|agx<¤o>=|o<&o>|n     |visa|age|xNoun|xAccusative
|vizagx<¤o>=|o<&o>|n     |face|xNoun|xAccusative

kiam
kiam     when

mi
mi     I

kovris
|kovr<¤i>=|i<&verb>|s     |cover|xPreterite|xIndicative

per
|per     |withHelpOf

kisoj
|kis<¤i>=|o<&o>|j     |kiss|xNoun|xPlural

tiujn
tiu|j|n     thatOne|xPlural|xAccusative

palajn
|pal<¤a>=|a<&a>|j|n     |pale|xAdjective|xPlural|xAccusative

sulkoplenajn
|sul<¤o>=|kopl<¤o>=|en<¤o>=|a<&a>|j|n
                            |?|?|yen|xAdjective|xPlural|xAccusative
|sulk<¤o>=|plen<¤a>=|a<&a>|j|n
                            |furrow|full|xAdjective|xPlural|xAccusative

vangojn
|vang<¤o>=|o<&o>|j|n     |cheek|xNoun|xPlural|xAccusative

enrigardadis
|en|rig<¤i>=|ard<¤i>=|ad|i<&verb>|s
|in|rig|glow|xContinued|xPreterite|xIndicative
|en|rigard<¤i>=|ad|i<&verb>|s
|in|look|xContinued|xPreterite|xIndicative

la
|la     the

bluan
|blu<¤a>=|a<&a>|n     |blue|xAdjective|xAccusative

okulon
|okul<¤o>=|o<&o>|n     |eye|xNoun|xAccusative
|ok&NumHyphOk|ul<¤o†bo>|o<&o>|n     |8_|xPerson|xNoun|xAccusative

en
|en     |in

kiu
kiu     who

vidigxis
|vid<¤i>=|igx<¤i&igx>|i<&verb>|s     |see|xBecome|xPreterite|xIndicative

tiom
tiom     thatMuch
```

```
da
|da     |ofQuantity

boneco
|bon<¤a>=|ec<¤o†bo>|o<&o>     |good|xAbstractQuality|xNoun

kaj
|kaj     and

amo
|am<¤i>=|o<&o>     |love|xNoun

longe
|long<¤a>=|e<&e>     |long|xAdverb

gxi
gxi     it

estas
|est<¤i>=|a<&verb>|s     |be|xPresent|xIndicative

kiam
kiam     when

min
mi|n     I|xAccusative

je
|je     |universalPreposition

lasta
|last<¤a>=|a<&a>     |last|xAdjective

fojo
|foj<¤o>=|o<&o>     |occasion|xNoun

benis
|ben<¤i>=|i<&verb>|s     |bless|xPreterite|xIndicative

sxiaj
sxi|a|j     she|possesiveA|xPlural

maljunaj
|mal<•verb>|jun<¤a>=|a<&a>|j     |xOpposite|young|xAdjective|xPlural

manoj
|man<¤o>=|o<&o>|j     |hand|xNoun|xPlural

se
|se     if

mi
mi     I

scius
|sci<¤i>=|us<&verb>     |know|xKonjunktive

majstre
|majstr<¤o>=|e<&e>     |maestro|xAdverb

per
|per     |withHelpOf
```

```
peniko
|penik<¤o>=|o<&o>     |?|xNoun

labori
|lab<¤o>=|or<¤o>=|i<&verb>     |?|gold|xInfinitive
|labor<¤i>=|i<&verb>     |labor|xInfinitive

mi
mi     I

vin
vi|n     you|xAccusative

glorus
|glor<¤i>=|us<&verb>     |glorify|xKonjunktive

kara
|kar<¤a>=|a<&a>     |dear|xAdjective

avineto
|av<¤o†bo†praFam>=|in<¤o†bo>|et|o<&o>
                              |grandfather|xFeminine|xDiminish|xNoun
|avi<¤o>=|net<¤a>=|o<&o>    |airplane|precise|xNoun

alie
|ali<¤a>=|e<&e>     |other|xAdverb

sed
|sed     but

mi
mi     I

ne
|ne     no

scias
|sci<¤i>=|a<&verb>|s     |know|xPresent|xIndicative

ne
|ne     no

scias
|sci<¤i>=|a<&verb>|s     |know|xPresent|xIndicative

kiel
kiel     how

tiu
tiu     thatOne

cxi
|cxi     near

skizo
|skiz<¤i>=|o<&o>     |sketch|xNoun

plume
|plum<¤o>=|e<&e>     |pen|xAdverb

desegnita
|de|seg<¤i>=|nit<¤o>=|a<&a>     |ofXFrom|saw|?|xAdjective
|desegn<¤i>=|i<&verb>|t<&part>|a<&a>
|design|xPreterite|xPassiveParticiple|xAdjective
```

```
al
|al<•o>     |to

iuj
iu|j     someOne|xPlural

ekplacxos
|ek|placx<¤o>=|o<&verb>|s     |xCommencement|?2*|xFuture|xIndicative
|ek|placx<¤i>=|o<&verb>|s     |xCommencement|please|xFuture|xIndicative
```

Appendix B.2 Words with more than one analysis

I have listed few most common or interesting examples from the list of words with more that one analysis. The list is obtained from the output of the program `filterBad`[71]. The full list is on the accompanying diskette.

```
ili
|il<¤o>|i<&verb>     |xTool|xInfinitive
ili     they
   fq:3265
aux
|aux     or
aux     ~interjection
   fq:988
avineto
|av<¤o?bo?praFam>=|in<¤o?bo>|et|o<&o>
                             |grandfather|xFeminine|xDiminish|xNoun
|avi<¤o>=|net<¤a>=|o<&o>    |airplane|precise|xNoun
   fq:843
sinjoro
|sin|jor<¤o>=|o<&o>     |xSelf|skid|xNoun
|sin<¤o>=|jor<¤o>=|o<&o>     |bosom|skid|xNoun
|sinjor<¤o>=|o<&o>     |gentleman|xNoun
   fq:702
soldato
|sol<¤o>=|dat<¤o>=|o<&o>     |solo|date|xNoun
|sol<¤a>=|dat<¤o>=|o<&o>     |alone|date|xNoun
|sold<¤o>=|a<&verb>|t<&part>|o<&o>
                      |?moneyPaiedToMercenary|xPresent|xPassPart|xNoun
|soldat<¤o>=|o<&o>     |soldier|xNoun
   fq:262
doktoro
|dok<¤o>=|tor<¤o>=|o<&o>     |dock|torus|xNoun
|dokt<¤a>=|or<¤o>=|o<&o>     |erudite|gold|xNoun
|doktor<¤o>=|o<&o>     |doctor|xNoun
   fq:262
rigardis
|rig<¤i>=|ard<¤i>=|i<&verb>|s     |rig|glow|xPreterite|xIndicative
|rigard<¤i>=|i<&verb>|s     |look|xPreterite|xIndicative
   fq:256
patrino
|pat<¤o>=|rin<¤o>=|o<&o>     |fryingPan|?|xNoun
|patr<¤o?bo?iUjCtr>=|in<¤o?bo>|o<&o>     |father|xFeminine|xNoun
   fq:250
barbara
|bar<¤i>=|bar<¤i>=|a<&a>     |obstruct|obstruct|xAdjective
|barb<¤o>=|ar<¤o>|a<&a>     |beard|xCollection|xAdjective
|barbar<¤o>=|a<&a>     |barbar|xAdjective
barbara     Barbara
   fq:207
subite
```

[71] See Appendix A.2.

```
|sub|i<&verb>|t<&part>|e<&e>     |under|xPreterite|xPassPart|xAdverb
|subit<¤a>=|e<&e>     |sudden|xAdverb
   fq:195
komprenas
|kom<¤o>=|pren<¤i>=|a<&verb>|s     |comma|take|xPresent|xIndicative
|kompren<¤i>=|a<&verb>|s     |understand|xPresent|xIndicative
   fq:149
vilagxo
|vil<¤o>=|agx<¤o>=|o<&o>     |shaggy|age|xNoun
|vilagx<¤o>=|o<&o>     |village|xNoun
fq:38
ene
|en|e<&e>     |in|xAdverb
|en<¤o>=|e<&e>     |yen|xAdverb
fq:30
vespermangxo
|vesp<¤o>=|er<¤o>|mangx<¤i>=|o<&o>     |wasp|xElement|eat|xNoun
|vesper<¤o>=|mangx<¤i>=|o<&o>     |evening|eat|xNoun
   fq:30
okulo
|okul<¤o>=|o<&o>     |eye|xNoun
|ok&NumHyphOk|ul<¤o?bo>|o<&o>     |8_|xPerson|xNoun
   fq:29
papero
|pap<¤o>=|er<¤o>|o<&o>     |pope|xElement|xNoun
|paper<¤o>=|o<&o>     |paper|xNoun
   fq:29
```

Appendix B.3 Unanalyzed words

I have listed unanalyzed words with frequency higher than 6 as obtained by the program `filterBad`[72]. The number in parentheses means frequency of the word in the corpus. Words that are proper names at first sight are marked with an asterisk. The full list is on the accompanying diskette.

The words containing *?* were in character set “Kamenicky” that is not recognized by program `toCorpus`[73].

rikitan (471*)
kefalín (466*)
saturnin (296*)
vane~k (268*)
vlahx (172*)
vilik (165*)
barunka (143*)
roy (114*)
viktorka (112*)
kristla (101*)
adélka (91*)
míla (87*)
giuseppe (85*)
smirnov (75*)
komteso (69)
jesxuo (68*)
iv-a (67)
pro?ek (64*)
nunenun (57)
jasánek (54*)
kefal (52)
venceslao (49*)
cxerník (45*)
halík (45*)
beyer (41*)
pr~emyslidoj (41*)
hamleto (40*)
hamácxek (36*)
nikeforo (35)
don (34)
padre (34)
kefalín-on (32*)
hamá (30*)
pro?ková (30*)
ii-a (29)
nepomuk (29*)
z~ernov (29*)
terezka (27*)
abraham (26*)
manc~inka (26*)
rosenkrantz (26*)
vilém (26*)
kudrna (25*)
be~tka (24*)
guildenstern (24*)
lot (24*)
pardubice (24*)
pr~emysl (24*)
lucio (23*)
frg (22)
otakar (22*)
staré (22*)
depost (21)
desur (21)
tamar (21)
be~lidlo (20*)
dale (20*)
janovice (20*)
manfred (20*)
sara (20*)
tábor (20*)
vlocxka (20*)
ferdinando (19*)

[72] See Appendix A.2.
[73] See Appendix A.1.

mar~enka (19*)
oliver (19*)
vohá (19*)
vor?a (19*)
hynek (18*)
ofiro (18)
antonín (17*)
euxpatoro (17)
halusxka (17)
luzacio (17*)
sodom (17)
antosx (16*)
ciml (16)
dinah (16)
friedel (16*)
grünfeld (16*)
htb-anoj (16)
janecxek (16*)
ralf (16*)
sir (16)
baksxi (15)
cilka (15*)
cxilpan (15)
far (15)
hamacxek (15*)
novotná (15*)
orel (15*)
pater (15)
ráb (15)
sultán (15*)
suza (15)
tome? (15*)
isahxar (14)
jeník (14*)
kovacx (14*)
mlle (14)
nikolao (14*)
romeo (14*)
sxtjetka (14*)
troník (14*)
bullio (13)
claire (13)
harlow (13*)
tiame (13)
viktorka'n (13*)
vilím (13*)
záruba (13*)
ahxilo (12*)
donh (12)
farell (12)
gt (12)
gxustatempe (12)
hirotaka (12*)
lt (12)
madame (12)
masaaki (12*)
roy-on (12*)
ruzyne~ (12*)
tyrl (12*)
an (11)
ankorauxfoje (11)
krist-infano (11)
macer (11)
orlík (11*)
sxt (11)
toník (11*)
ve~ra (11*)
vonjavka (11)
afirmis (10)
ekveis (10)
her~man (10*)
ildefonso (10)
jerry (10*)
johanka (10*)
kahoun (10*)
luksemburgia (10)
maniko (10)
pr~emyslida (10)
ruml (10*)
strobo (10)
terebová (10*)
vampera (10)
vid-al-vide (10)
xiv-a (10)
andula (9*)
c~ervená (9*)
darfis (9)
erik (9*)
hilla (9)
hostiva (9*)
ippolito (9)
janek (9*)
ka (9)
kladsko (9*)
kohl (9*)
korsor (9)
krkono?e (9*)
kunte (9)
ludek-on (9*)
na (9)
pariso (9*)
rapid (9)
sláma (9*)
sunsubiro (9)
svatopluk (9*)
vekril (9)
'sur (8)
cavaliere (8)
dubský (8*)
hohxman (8)
hunnoj (8)
hus (8*)
i-a (8)
ii (8)
iii (8)
ijk (8)
jakub (8*)
janýsek (8*)
kohn (8*)
lidu?ka (8*)
madlenka (8*)
nahxumo (8)
neol (8)
nikomahxo (8)
ole?nice (8*)
oplt (8)
papanastasias (8)
posteuloj (8)
praha-n (8*)
sanktvenceslaa (8*)
senpere (8)
stone (8)
turyn~ (8*)
vondrou?ek (8*)
zuzanka (8*)
barras (7)
blanjo (7)
bohdalová (7*)
br~etislav (7*)
dobru?ka (7*)
fialho (7)
fousková (7*)
german-franca (7)
giovanni (7*)
henriko (7*)
hradová (7*)
hu°ra (7*)
in (7)
jaroslav (7*)
jesse (7)
joanna (7*)
kadera (7)
le (7)
lear (7)
luksemburgiidoj (7)
mandros (7)
manikoj (7)
marbach (7)
psxa (7)
sajner (7)
slujxka (7)
sodomon (7)
susxice (7*)
trojanoj (7)
trojo (7*)
vamber~ice (7)
vilik-on (7*)
vorsxa (7*)
vosteto (7)
voston (7)
x-a (7)
xi-a (7)

Appendix C Two-level rules

Here, I list all two-level rules. The description for rules was provided in the chapter 4 Implementation. For description of the formalism, see chapter 4.1 Two-level morphology. I have listed also tables for finite state automata.

```
ALPHABET
       b c d f g h j k l m n p q r s t v w x y z a e i o u
       cx gx hx jx sx ux
       á é í ó ú ý š ß ~ u°
       ' - +
       =             ;end of the classicla root
       |             ;beginnig of the new morpheme
       /             ;end of the word in coordinative composites
       < >           ;morphological features (for better readibility)
       ¤o ¤a ¤i ¤e ¤w             ;categories of the root
       © ®           ;possibility/necesity to include o before vocal
       &NumHyphen &NumHyphOk
       &part         ;to forbid multiply application of particple suf.
       •root •neverRoot
       ^Err^
;Infl
       &o      •o
       &a      •a
       &e      •e
       &verb   •verb       ;-i, -us, -u, -a/i/o|s/nt/t
;Prefixes&Prefixoids
       &bo †bo
       &ge
       &praFam †praFam
       &praPri
;Suffixiodes
       &ig     •ig
       &igx    •igx
       &iUjBox
       &iUjCtr †iUjCtr
       &iSci   †iSci

NULL 0
ANY @
BOUNDARY #

;consonants
SUBSET C  b c cx d f g gx h hx j jx k l m n p q r s sx t v z
;vocals
SUBSET V   a e i o u ŭ
;features
SUBSET X < > © ¤o ¤a ¤i ¤e ¤w &NumHyphen &NumHyphOk &bo †bo &ge &iSci
†iSci †iUjCtr &iUjCtr &o •o &a •a &e •e &verb •verb
SUBSET B / |              ;begining of a morpheme, of 2nd word in
coordinative composite


RULE "defaults" 1 31
    b c d f g h j k l m n p q r s t v w x y z a e i o u ' + | - @
    b c d f g h j k l m n p q r s t v w x y z a e i o u ' 0 0 - @
 1: 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

```
RULE "defaults" 1 7
    cx gx hx jx sx ux @
    cx gx hx jx sx ux @
 1: 1  1  1  1  1  1  1

RULE "defaults" 1 10
       á é í ó ú ý š ~ u° @
       á é í ó ú ý š ~ u° @
 1: 1 1 1 1 1 1 1 1 1  1

RULE "defaults" 1 16
   + = | / < > ¤o ¤a ¤i ¤e &NumHyphen &NumHyphOk &part © ® @ ;© © ® @
   0 0 0 0 0 0 0  0  0  0  0          0          0     0 o @ ;0 o o @
1: 1 1 1 1 1 1 1  1  1  1  1          1          1     1 1 1 ;1 1 1 1

RULE "defaults" 1 10
    •root &o •o &a •a &e •e &verb •verb @
    0     0  0  0  0  0  0  0     0     @
 1: 1     1  1  1  1  1  1  1     1     1

RULE "defaults" 1 7
    &bo †bo &ge &praPri &praFam †praFam @
    0   0   0   0       0       0       @
 1: 1   1   1   1       1       1       1

RULE "defaults" 1 9
    &ig •ig &igx •igx &iUjCtr †iUjCtr &iSci †iSci @
    0   0   0    0    0       0       0     0     @
 1: 1   1   1    1    1       1       1     1     1

;1 means (¬|:0)* |:0 (¬|:0)* - anything but |:0 only one - next
morpheme

;------------------------------------------------------------
;Misspellings

;aux is sometimes by mistake written as au
RULE " ux:u => [a|e|o]__ @ 0:^Err^#" 4 7
      a  e o ux ^Err^ # @
      a  e o u  0     # @
 1:   2  2 2 0  0     1 1
 2:   2  2 2 3  0     2 1
 3:   3  3 3 3  4     0 3
 4:   0  0 0 0  0     1 0


;------------------------------------------------------------

;hx after r can be replaced by
RULE " hx:k => r __" 2 3

      r  hx @
      r  k  @
 1:   2  0  1
 2:   2  1  1
```

```
;One alternative of inserting an o (only between two consonants)
;RULE "=:o => C __ |:0 C" 5 5
;    = | C X @
;    o 0 C 0 @
; 1: 0 1 2 1 1  ;std
; 2: 3 1 2 2 1  ;C
; 3: 0 4 2 3 1  ;C =:o
; 4: 0 0 5 4 0  ;C =:o |:0
; 5: 3 0 2 5 1  ;C =:o |:0 C

RULE "=:o => __ |:0 (¬|:0)* [=:o | =:0]" 3 4
    = = | @
    o 0 0 @
 1: 2 1 1 1  ;std
 2: 3 0 3 0  ;=:o
 3: 3 1 0 3  ;=:o |:0 ..

RULE " &NumHyphen => ___ (¬/:0)* &NumHyphOk" 3 5
      &NumHyphen &NumHyphOk / # @
      0          0         0 # @
 1:   2          1         1 1 1
 2:   2          3         0 0 2
 3:   3          3         1 1 3

;--- bo & ge ------------------------------------------_

;bo requires allowed root, between bo and that root can be only ge
RULE "&bo => __ 1 (&ge 1) (†bo) " 3 5
      &bo &ge | †bo @
       0   0  0 0   @
 1:    3   1  1 1   1           ;std
 2:    0   3  0 1   2           ;&bo (prev morpheme)
 3:    0   0  2 1   3           ;&bo (this morpheme)

;bo|ge ok, ge|bo ko
RULE "&ge /<= __ 1 &bo " 3 4
      &ge | &bo @
       0  0  0  @
 1:    3  1  1  1         ;std
 2:    0  1  0  2         ;&ge previous morpheme
 3:    0  2  0  3         ;&ge this morpheme

;ge cannot be chained (*ge|ge|patro)
RULE "&ge /<= __ 1 &ge " 3 3

       &ge |  @
       0   0  @
 1:    3   1  1           ;std
 2:    0   1  2           ;&ge (previous morpheme)
 3:    0   2  3      ;&ge (current morpheme)
```

```
;--- pra ------------------------------------------------

;Pra for Families
RULE "&praFam => __ 1 [ †praFam | &praFam ]" 3 4

      &praFam |  †praFam @
       0      0  0       @
 1:    3      1  1       1       ;std
 2:    3      0  1       2       ;&praFam (previous morpheme)
 3:    3      2  1       3       ;&praFam (current morpheme)

;Roots that can get praFam, cannot get praPri, &praPri and &praFam
cannot be combined
RULE "&praPri /<= __ 1 [ †praFam | &praFam ]" 3 5

       &praPri | †praFam  &praFam @
       0       0 0        0       @
 1:    3       1 1        1       1     ;std
 2:    3       1 0        0       2     ;&praFam (previous morpheme)
 3:    3       2 0        0       3     ;&praFam (current morpheme)

;PraPri cannot be chained
RULE "&praPri /<= __ 1 &praPri " 3 3

       &praPri |  @
       0       0  @
 1:    3       1  1        ;std
 2:    0       1  2        ;&praFam (previous morpheme)
 3:    0       2  3        ;&praFam (current morpheme)

;--- iSci, iUjCtr ----------------------------------------

;Science suffix i
RULE "&iSci => †iSci 1 * __" 3 4

      †iSci |  &iSci @
       0    0  0     @
 1:    3    1  0     1          ;std
 2:    3    1  1     2          ;†iSci (suffix iSci can be applied)
 3:    3    2  1     3

;Inhabitant can not be put into a box
RULE "&iUjBox /<= †iUjCtr 1 __" 3 4

      †iUjCtr |  &iUjBox @
       0      0  0       @
 1:    3      1  1       1       ;std
 2:    3      1  0       2       ;†iUjCtr (previous morpheme)
 3:    3      2  0       3       ;†iUjCtr (current morpheme)

;Country suffixes i or uj
RULE "&iUjCtr => †iUjCtr 1 __" 3 4

      †iUjCtr |  &iUjCtr @
       0      0  0       @
 1:    3      1  0       1       ;std
 2:    3      1  1       2       ;†iUjCtr (previous morpheme)
 3:    3      2  1       3       ;†iUjCtr (current morpheme)
```

```
;--- Participles ------------------------------------------

;There cannot be multiple participle suffix in the word (can be in
two parts of coordinate composites)
RULE "&part /<= &part (¬/)* __ " 2 3

       &part / @
       0     0 @
 1:    2     1 1          ;std
 2:    0     1 2          ;&praFam (previous morpheme)

;--- Forbiding some infl morphemes ------------------------

;Second | forgets the forgiving feature (first is at the begining of
the following morphem)
;*al<•o>|o<&o>            (| does not forget => impossible)
; al<•o>|ec|o<&o>   (2nd | forgets => possible)
;For explication of 1 see up

RULE "&o /<= •o 1 __ " 3 4
      •o |  &o  @
      0  0  0   @
 1:   3  1  1   1          ;std (no •o in this or prev. morpheme)
 2:   3  1  0   2          ;•o (previous morpheme)
 3:   3  2  0   3          ;•o (current morpheme)
```

The next five tables for finite state automata are the same as the one for a noun.

```
RULE "&a /<= •a 1 __ " 3 4

RULE "&e /<= •e 1 __ " 3 4

RULE "&verb /<= •verb 1 __ " 3 4

RULE "&ig /<= •ig 1 __ " 3 4

RULE "&igx /<= •igx 1 __ " 3 4

RULE "=:0 /<= •root 1 __ " 3 4
      •root | = @
      0     0 0 @
 1:   3     1 1 1
 2:   3     1 0 2
 3:   3     2 0 3

RULE "=:0 /<= •neverRoot (¬/)* __ " 2 4
      •neverRoot = / @
      0          0 0 @
 1:   2          1 1 1
 2:   2          0 1 2
```