REZEKNE ACADEMY OF TECHNOLOGIES
FACULTY OF ENGINEERING

ANSIS ATAOLS BĒRZIŅŠ

AUTOMATED COMPARISON OF NATURAL LANGUAGES

SYNOPSIS OF THE DISSERTATION
submitted in fulfillment of the requirements for the award
of a **degree of Doctor of Engineering** (***PhD.***)
in the field of Information technology

Rositten, MMXX

The research was carried out at:

Rezekne Academy of Technologies
Faculty of Engineering,

University of Latvia,
Faculty of Physics and Mathematics

from year 2003 to 2020.

Dissertation director:

Artis Teilāns
Dr. sc. ing.,
professor at the Department of Computer Science and Mathematics,
Faculty of Engineering,
Rezekne Academy of Technologies

Partial co-advisors:

Pīters Grabusts
Dr. sc. ing.,
professor at the Department of Computer Science and Mathematics,
Faculty of Engineering,
Rezekne Academy of Technologies

Mats Blomberg
retired researcher at the Division of Speech, Music and Hearing (TMH),
KTH Royal Institute of Technology in Stockholm

Lukáš Burget
doc. Ing., Ph. D.,
associate professor at the Department of Computer Graphics and Multimedia,
Faculty of Information Technology,
Brno University of Technology

NB This is just a synopsis! In case you really want to understand the topic completely, it's recommended to read the full DISSERTATION.

Submitted for defence to the Promotional Council in the field of Information Technology of the Rezekne Academy of Technologies in year 2020.

Content

I Work with text data in phonetic transcription

II Work with speech data in sound recordings

III Verification

[1] The original numbering of the dissertation is kept in the synopsis.

Preface

We have outlined our path to the topic of the work and its solution in the introduction, while the formal side of our message is presented here – in the preface.

It should be noted that, although the dissertation is presented for defense in the field of engineering, the study itself is completely interdisciplinary and could be defended in computer science, applied mathematics or linguistics as well. The classification and categorisation of such cross-sectoral works generally depends on a local academic tradition.

The **specificity of the work** is the adaptation of methods used in other fields of computational linguistics[2] to the field of linguometry or dialectometry.

The **topicality of the work** is determined by its timelessness: speculations on proximity of languages and resulting political consequences have been actual throughout periods of development of humanity, during which people felt a national self-confidence and its linkage with a language. At the beginning of the XX century, with the widespread formation of nation-states, this issue has become particularly topical, and in the XXI century its topicality has not decreased.

The **goal of the** present **work** is development of methods that would allow to quantify a degree of proximity of natural languages. In a case of successful implementation this assessment should meet certain requirements:

- it must be a metric, ie., fulfill the axioms of symmetry, triangle and identity or non-degeneracy;
- it must comply with socially intuitive notions of language proximity (which in most cases correspond to linguistically analytical views too);
- it must be applicable to as wide a range of languages as possible, preferably to all natural human languages;
- the cost of preparation of input data must be as low as possible, especially in the meaning of manual work.

The following **tasks** were set and solved in order to achieve the goal:

- finding and selecting input data formats, defining data preparation requirements;
- selection and adaptation of algorithms of methods;
- collection and preparation of pilot data;
- testing and evaluation of methods.

The dissertation **consists** of an introduction and three parts, the first dealing with phonetic transcriptions, the second dealing with speech recordings and the third one dealing with verification of all the results. Two first parts are divided into several chapters, most chapters – in several sub-chapters.

The **main results** of the scientific work:

- appropriate input data types have been selected and technical requirements they have to meet have been defined;
- a dialect speech corpus has been manually collected and created;
- six independent, different methods of distance measurement have been developed:
    - method of n-grams for integral, non-parallel texts (see Chapter 3);
    - method of edit distances for parallel sets of words (see Chapter 5);

[2] Off course, we see the computational linguistics as a superset that encompasses natural language processing and language technology.

- method of phoneme recognisers (see Chapter 8);
- method of hidden Markov models for full-size recordings (see Chapter 9);
- method of i-vectors for full-size recordings (see Chapter 10);
- method of Gaussian mixture models for full-size recordings (see Chapter 11);

- software to implement these methods has been adapted and developed;
- verification and comparison of the results has been carried out;
- significant scientific intermediate results have been obtained, such as newly developed space of phonems (see Chapter 4) and hierarchical choices metrics (see Chapter 12).

The work is **scientifically innovative** because:

- although the issue of proximity of languages has flurried people's minds as such already for a long time, and in recent decades there have been scientific attempts to solve it with help of a computer too, we have offered an unprecedented approach – as a basis of language to take speech rather than writing, thereby significantly widening a range of languages considered, and also making methods more objective, independent of conditional ortographic systems;
- new applications have been found for methods used in other areas of computational linguistics (text classification, speech recognition, etc.) , e.g., a completely new approach is to generate statistical models of speech recognition from full informants' recordings rather than specific words, thus creating models that represent a language as a whole;
- a new concept – the space of phonemes – has been introduced and described defining dimensions that coordinate all phonemes of Latvian and Latgalian dialects, defining anatomically and phonetically based metrics (so far no one has assembled both vowels and consonants in a same space with distances that comply with intuitive understanding of phonemes’ proximity);
- not only expected, but also unexpected results of experiments have been obtained, which can be scientifically explained and substantiated (see., eg., the conclusions at the end of Chapter 5);
- a new metrics – hierarchical choices metrics – has been introduced, which allows a numerical comparison of binary trees resulting from hierarchical clustering, axioms of metrics have been proven.

The **results** of the work **practically implemented** as:

- PERL CGI scripts for calculating and graphical displaying a table of distances of the space of phonemes;
- PERL script for applicating the method of n-grams;
- PERL scripts for applicating the methods of edit distances (Levenshtein and Wagner-Fischer);
- PERL scripts for applicating of the phoneme recognisers *PhnRec* and *Sphinx* and analysing the results;
- PERL script for applicating hidden Markov models using the HTK package;
- PERL script for applicating the method of i-vectors;
- PERL scripts for categorising of the results;
- Python script for applicating cosine similarity;
- Python script for applicating Kullback-Leibler divergence;
- MatLab script for generating Gaussian distribution models of full-length recordings without MAP adaptation;
- MatLab script for generating Gaussian distribution models with MAP adaptation;
- MatLab script for calculating various distances (Euclid, L2, Kullback-Leibler, Jordan) between super-vectors converted from matrices;
- MatLab script for calculating special Mahalanobis distance between Gaussian distributions;

- MatLab script for calculating special Euclidean distance between Gaussian distributions;
- PERL script for calculating hierarchical choices metrics;
- and other smaller software units.

During the period 2004-2019 **results** of the scientific work were **discussed** at:

- international scientific **conferences**:
    - «Корпусная лингвистика», Saint Petersburg, Russia, 2004;
    - «Identification des langues et des variétés dialectales par les humains et par les machines», Paris, France, 2004;
    - «Диалог», Moscow district, Russia, 2006;
    - J. Endzelin’s conference, Riga, Latvia, 2007;
    - «Megaling», Kiev, Ukraine, 2009;
    - „ქართული ენა და თანამედროვე ტექნოლოგიები“, Tbilis, Georgia, 2011;
    - «Диалог», Moscow, Russia, 2016;
    - «Artificial Intelligence and Natural Language», Tartu, Estonia, 2019;
    - «Открытая конференция ИСП РАН им. В.П. Иванникова», Moscow, Russia, 2019;
- **doctoral schools**:
    - PhD school winter session, Distance Education Center, Riga Technical University, Ives par., 2011;
    - seminar-competition, "Research Slam"Riga Technical University Doctoral School, Riga, 2015;
- **seminars**:
    - seminars of Artificial Intelligence Laboratory, Institute of Mathematics and Informatics, University of Latvia, 2004-2007;
    - doctoral seminar of Department of Computer Science, Faculty of Physics and Mathematics, University of Latvia, 2008;
    - seminars of Distance Education Center, Riga Technical University, 2009-2013;
    - seminars of Faculty of Engineering, Rezekne Higher Education Institution, 2014-2016;
- during **internships**:
    - Division of Speech, Music and Hearing, Stockholm,Royal Institute of Technology, Sweden, 2008;
    - Speech Processing Research Group, Department of Computer Graphics and Multimedia, Faculty of Information Technology, Technical University of Brno, Moravia, Czech Republic, 2015;
    - Lithuanian Language Institute, Vilnius, Lithuania, 2015;
    - Speech Processing Laboratory, Faculty of Engineering, National Autonomous University of Mexico, Mexico, 2016.

We have summarized results of the dissertation in **9** scientific **publications**:

¤ Берзиньш А.У. Сравнение балтийских языков методом n-грамм // Труды международной коференции «Корпусная лингвистика - 2004». СПб.: Издательство С.-Петербургского университета, 2004.

¤ Berzinch A.A. La comparaison de typologie traditionnelle et de typologie phonolexique, basée sur la méthode des n-grammes, dans les dialectes baltes // Identification des langues et des variétés dialectales par les humains et par les machines. Paris: École National Supérieure des Télécommunications, 2004.

¤ Берзинь А.У. Измерение фономорфолексического расстояния между латышскими наречиями путём применения расстояния Вагнера-Фишера // Труды международной конференции «Диалог 2006». М.: Издательство РГГУ, 2006.

¤ Bērziņš A.A., Grigorjevs J. Latviešu izloksnēs sastopamo fonēmu telpa // Linguistica Lettica XVIII. R.: Latviešu valodas institūts, 2008.

¤ Берзинь А. Возможности применения статистических методов распознавания речи для определения близости языков // Прикладна лінгвістика та лінгвістичні технології, Megaling-2009. Київ: «Довіра», 2009.

¤ ბერზინი ა. ინფორმაციის მოპოვების პრინციპები ფონოგრამების ავტომატური ანალიზისთვის / Принципы сбора информации для автоматизированного анализа фонограмм // ქართული ენა და თანამედროვე ტექნოლოგიები - 2011. თბილისი: „მერიდიანი“, 2011.

¤ Берзинь А.У. Применение распознавателей фонем для автоматического определения уровня близости языков // Труды международной конференции «Диалог 2016». М., 2016.

¤ Bērziņš A.A. Usage of HMM-based Speech Recognition Methods for Automated Determination of a Similarity Level between Languages // AINL Proceedings. «Springer», 2019. *Indexed in* ***Scopus*** *and* ***Web of Science****.*

¤ Берзинь А.У. Применение i-векторов для автоматизированного определения уровня близости языков // Труды Института системного программирования РАН. Том 31, № 5. М.: ИСП РАН, 2019.
Indexed in ***Российский индекс научного цитирования*** *(РИНЦ) and* ***КиберЛенинка****.*

1 of dissertation results has been **submitted** for publication:

¤ Berziñch A.A., Chavarría Amezcua M.A. El espacio de los alófonos del español. Presentado para publicación en «Lengua y Habla», 2020.

We have also published **8** scientific publications in other fields of science, including 1 in machine translation, 2 in computer lexicography, 3 in comparative linguistics, 1 in ethnomusicology and 1 in comparative folkloristics.

Another **1** scientific publication in other field – terminology – has been submitted for publication.

Introduction

There are different hypotheses when human beings started to speak. For example, Australopithecines have already walked upright, bringing their breathing and speech apparatus to a more suitable, comfortable position. Therefore, researchers assume that elemental speech has already been spoken, even though a communication of today's anthropoid apes is not considered a speech, because it’s just a communication by separate cries, but a prerequisite for speech is the presence of thought. There are also signs of anatomical development that might be related directly to speaking, for example, Acheuleans had it. It is believed that Neanderthal society has already been talking. In any case, speech, as a means of communication, probably was already available from the very beginning of the modern human or *Homo sapiens sapiens*.

According to the linguistics classicist F. Saussure, speech is a form of individual expression, whereas language is already a system of conventions that can be used by an individual who is familiar with it. So in fact, on the matter of our interest, we cannot separate a speech from a language. We should assume that they go almost hand in hand: it will not make sense for a human to speak if no one understands him, that is, to make a speech act that does not conform to any public language convention.

There are various hypotheses about the origin of languages. For example, the theory of monogenesis says that all languages descended over time from one source language. If *Homo sapiens sapiens* really came from one place in Africa, where a mutation occurred, then this assumption would be logical. But there are other options too. Anyway, it is clear that as soon as linguistic differences emerged, so did the problem of proximity of languages: closer ones was understandable, further ones – was not, and people classified them somehow. So this problem, albeit on an intuitive level, has been a matter in people's minds for thousands of years.

The first known publications on proximity of languages in Europe appeared in the XVI century, but at a scientific level this issue was raised by William Jones in the XVIII century.

Historically, comparative linguistics only tends to compare languages of a related origin, i.e. some that has a common history. While this approach is natural and seeks answers to most important questions of language proximity, we believe it is already outdated and expandable. Many non-congenial languages of origin have been cohesive for years, influencing each other, so proximity or distance is not only a comparison of placement of sheets in the tree of origin, but an assessment of differences in the overall contribution of all phases of language development. For example, knowing the influence of Livonian in Latvian, the issue of closeness of Estonian, Finnish and other Finno-Ugric languages to Latvian is not so pointless, although in terms of origin they are not related (of course, the influence of German on Latvian and Estonian and the influence of Russian on Latvian and Finnish only increase this interest). In other words, from an actual point of view, it is not only a historical-comparative, but also a purely comparative assessment that is important.

In the 200 years, historically-comparative linguistics has fulfilled its mission - most of the known languages of the world are genealogically categorised (though, of course, as in any sector, potential tasks to go more deeply will never be lacking). However, it does not provide a numerical assessment of this proximity, although social processes sometimes require such a thing.

It is not a coincidence that the expression, published in 1945 by the Goldingen[3]-born Meyer (Max) Weinreich, that a language is a dialect plus an army and a navy (he heard it from a listener of one of his lectures), can now be called winged and well known.This definition struck a nail in the head, describing influence of political processes on linguistic issues. Should it be? Probably – no, science, including linguistics, should be objective and therefore independent of political orders. Until now, however, despite the rapid development of science and computing

[3] A city in Courland.

capacity since 1945, a mechanism that would allow objective measurements to refute this kind of political speculations has not yet been created.

We felt the need for it in the early 2000s, before starting our doctorate. It also determined the choice of the topic of the dissertation. Later, we encountered Jahontov's publication (1980), which, in terms of justification, almost completely matched our motivation, but the scales offered by that author for any language and the two languages of one nation do not give a numerical rating, but an orderly assessment of 5 and 8 degrees. In other words, when differences in proximity of language pairs are large enough, we can find that one pair is further apart than the other one, but as soon as they enter into one "proximity group", there is no a finer weighing available.

Our attention also attracted J. Tambovtsev's work. He calculates a phonotypological distance between Basque and some other languages of different groups using Euclidean distance in an eight-dimensional space, whose dimensions are occurence frequencies of various consonants. And although the results of Tambovcev's experiments are very interesting, unfortunately we are not satisfied with them, because the author works only with official, literary languages, not with speech or phonetic transcription. Sources (books) written in a certain orthographic tradition he converts to an expected sound (e.g., a phonetic transcription of it) according to some formal spelling rules. There is no doubt that both the ortographical transcription and the subsequent transformation to phonetic transcription are conditional and, when viewed strictly, produce an artificial language, different from the real one. Comparing several languages obtained in the same way (even a strange one) is not incorrect; however, as we are interested in languages without a writing tradition too, we should look for other ways.

The fact that a numerical estimation would be the best proof American scientist M. Swadesh realised even in the mid-1950s. For glottochronology, the author has suggested to use a common vocabulary of language pairs (specifically selected to represent direct origin rather than later contacts) to calculate their segregation times in centuries (i.e., the number of centuries before the languages were separated) and to determine language's division according to this numerical criterion (i.e., how close (old) relatives they are). The author mainly deals with CVC (consonant-vowel-consonant) group coincidences and correspondences, and describes different methods for different situations (depending on language characteristics and relationships). However, such an assessment is only applicable to languages of the same superfamily, and only lexicon is taken into account, that is insufficient. And even though this estimation is, looking purely mathematically, a distance, i.e., corresponds to the axioms of metrics, the conditionality of its acquisition mechanism and the size of the scale step give reason to doubt whether it would be applicable in our case.

In our view, not only vocabulary, but also phonetics, morphology and syntax should be taken into account, since they also characterize proximity of a pair of languages. Clearly, in such a setting, it makes sense to work only with data that contains that information too, and is fixed, language-independent, universal. Therefore texts in orthographic transcription are completely unusable because they are conditional and ambiguous – one language can have several orthographic traditions and it's not possible clearly obtain a pronunciation from such a transcription. The conclusion is that either texts in phonetic transcription (with same universal transcription system being used for all languages) or speech recordings should be used. Thus, we put forward a **hypothesis** that during the development of the dissertation will try to examine: both phonetic transcription of speech and recordings of speech as input data are sufficient to derive a numerical estimation of the degree of proximity of languages by statistical methods.

Nowadays, the method of glottochronology is considered outdated, and phylogenetics are used instead, using as input data glottochronological vocabulary lists defined by Swadesh, supplemented or even not supplemented by lexical units relevant to the study in question. Developmental theory and trees were introduced by scientists already in the XIXth century, the name "phylogenetics" appeared in the 20s of XXth century, but the formation of phylogenetic trees for creature species was formalized in 1950 by German entomologist Wilhelm Henich, in

1965 he published his idea in English too. It turned out that the proposed method was very versatile and was gradually being used in other sciences and disciplines, such as genetics (to compare DNA of molecules) and, as may be understood from our interest, in linguistics. Linguistic phylogenetics of Indo-European languages was intensively used in the 50-60s of XXth century, since 70s additional statistical methods have been appeared too. Statistical phylogenetics is used even in very thorough and recent publications, which, for example, prove one of the two hypotheses of Indo-European languages' origin by means of phylogenetics. But the disadvantage of the method, as in glotochronology, is that it is applicable only to languages with a common origin.

The distances defined in this dissertation, of course, must not contradict people's intuitive understanding of proximity of and distance between languages. Every speaker of a language has an idea of what languages are closer to his or her language, what – are further, what he can understand almost entirely, what – partly and what – just a few words or forms. Of course, a person will have such a vision only about some languages: mainly languages of neighbouring nations and about large, international languages. Mostly this understanding reflects the actual situation in languages, without attempting to separate relationship of origin from later developments of cohabiting, although, of course, nation's historical memory may also play a role. Thus, this intuitive understanding is also largely in line with analytical language classifications: both the genealogical and typological.

The dissertation is arranged chronologically – as the topic developed, we were looking for methods that could help to solve the problem. Naturally, we first caught up with phonetic transcription (in fact, not just for texts, because one method was designed for parallel vocabulary), since it is more visible, then switched to recordings, whose analysis is, of course, more complicated, but the task was relieved by the fact that we used software packages already available for statistical analysis of speech.

Method of n-Grams for Integral, Non-Parallel Texts

In 1994, William Canvar and John Trenkle proposed to use n-gram frequency lists to categorise texts. The main idea of the method lies in the fact that, according to Zipf’s law, a set of words (or, as in our case, a string of graphemes or phonemes) can be ordered by their frequency of use. Canvar and Trenkle suggested to compile n-gram frequency lists for different texts and determine the category to which the input text belongs by comparing those lists with the n-gram frequency list of an input text you wish to categorise. This way the language, encoding and even topic of a text can be successfully automatically detected. What does this have to do with our tasks? The fact is, for the sake of comparing n-gram lists, a number representing a degree of similarity between the texts is calculated. Therefore, a question naturally arose: can we use this method to determine a degree of language similarity? Obviously, yes, but of course, just in case of usage of same phonetic transcription for all the languages compared: since different transcriptions and even different encodings, which are helpful for tasks of categorisation, in this case are damaging, even makes it meaningless. It should also be noted that requirements for the categorisation task are significantly weaker than in the case of distance calculation task, so we can foresee that we will have to work with more texts to obtain sufficiently accurate results.

The length of n-grams used (i.e., n) was chosen empirically. We set n=1..5, namely from unigrams to quintagrams. Of course, what n-grams will be represented in frequency lists also depends on the maximal size of a list (the number of entries in it): it should be defined in such a way that the list represents as much as possible the text on which it was created, but at the same time in all the texts there are enough of different n-grams to fulfil the list. In our case, we defined N=400 and, in order to keep the conditions the same throughout the experiments, we left the constant unchanged, which is inherently incorrect: this constant should be adapted to conditions of a particular experiment, most often to the amount of data, i.e., of input text.

In this experiment, using manually prepared texts in phonetic transcription (so they were relatively small), our lists contained all 5 sizes of n-grams allowed, thus the incidence of n-grams at the back of the lists was 1, i.e., N was theoretically close to the maximum possible, but practically even a little too large. It also determined that the distances are numerically large, as there are many rare n-grams that are less likely to be encountered in other languages. In our Chapter 8 experiment, which also used the n-gram method, but used auto-generated texts (hence large ones) for input data, our lists only contained unigrams, bigrams, and trigrams (since quadragrams and quintagrams are less common), n-gram frequency at the backs of the lists ranged from tens to a few hundred, so N was too small.

The known standards of digital storage of phonetic transcription (various local, Unicode, SAMPA, IPA/ASCII) do not satisfy our needs because they either do not contain all sounds, either they have a variable number of characters to represent one phoneme (i.e., they can be bytes, but can be byte pairs). It would be more convenient for us to work with single-size text that describes each sound with the same amount of information units. Since one-byte (256 permutations) is not enough to describe all human speech sounds, the most convenient way, apparently, would be to use two bytes (65536 permutations). Of course, even a finer division is possible, for example, if we consider not only phonemes but also their allophones as different sounds. However, we do not see benefits of this, but a proportion of transcribing errors and individual perception’s errors will certainly increase. Moreover, since existing phonetic transcription systems do not display such nuances, it will not be possible to use texts that have already been transcribed.

Obviously, in the future, it would be desirable to develop special Unicode fonts as well as an editor and converters for such a transcription system. For the time being, we worked with a two-byte pseudocode for working with speech transcriptions of Baltic languages: we described each phoneme we encountered with two ASCII characters, according to a self-defined scheme.

At the time of the experiment a corpus of texts of Baltic languages in phonetic transcription was not available so we had to input some dialectal texts by manual typing from books. Since we only did this for the sake of the experiment, we entered a small number of small-size texts: 4 samples of various Latvian subdialects (3 – Latgalian and 1 – Couronian) ranging from 500 to 1000 characters. We then adapted the PERL program *TextCat* (written in 1994 by Gertjan van Noord, inspired by Canvar and Trenkle) to work with two-byte, not one-byte, characters, i.e., n-grams were 2n bytes long instead of n as in the original program.

Initially, we obtained unsymmetrical (albeit plausible in terms of degree of proximity) results. As we knew they had to be symmetric, we started looking for an error and found two in the *TextCat* algorithm – in the original code of it – that led to this asymmetry. After correction, we obtained symmetric results.

Despite the small amount of data entered, the results of our experiment fully live up to our expectations. Two neighbouring North Latgalian subdialects – Baļtinova and Škylbani – are closest to each other, the next closest is Nierza, and the only Couronian subdialect turns out to be the furthest. The speech of Baļtinova appears to be the closest of Nierza, which is justified both geographically and linguistically, but the furthest, of course, is Džūkste. The closest to Džūkste is Škylbani, which is not surprising, because North Latgalian subdialects have more in common with non-Latgalian ones both lexically and from the side of vocal presentation of certain grammatical forms, but farest – Nierza, which is indeed a “deeper” Latgalian both lexically and morphologically. Baļtinova subdialect, on the other hand, turns out to be slightly closer to the Nierza than Škylbani; this is also in line with analytical understanding, as Baļtinova has a greater influence of Kuorsova (which, in turn, is closer to the Nierza subdialect) than Škylbani has.

Thus, with very small texts and a binary comparison procedure, we have achieved reasonably (even very) good results. Therefore, there is reason to believe that this method can be used to determine a degree of proximity of languages, especially with larger corpora of texts in phonetic transcription. Perhaps even better results could be obtained by comparing rows of frequency lists with some other distance, which would describe not only the coincidence, but also the degree of similarity, for example, Levenshtein or Wagner-Fischer distance.

We published the results of this experiment in 2004. Many years later, in 2016, we learned from Indian colleagues that in 2011 they conducted a similar experiment with Indian languages. They did not work with transcriptions of real speech, but emulated them from orthographic texts, converting them into a common phonetic transcription – it means their method is applicable to real phonetic transcriptions too. Unlike our method, they construct n-gram probability models and then calculate the L_1 and L_2 norms, as well as Kullback-Leibler and Rao divergences. In the future, it would be interesting to try their method on our data and compare the results.

Space of Phonemes. Concept. Realisation in Latvian Dialects

We came across a need to place all Latvian phonemes (both Baltian[4] and Latgalian) in one multidimensional coordinate system, or space, when we decided to measure distances between phonetic transcriptions using Wagner-Fischer distance (this is described in the next chapter): a distance between any two graphemes (=phonemes, since it was a phonetic transcription) needed to be defined and it was clear that such a distance must be phonetically based.

Edit distances, including Wagner-Fischer distance, are used to quantify a level of dissimilarity of character strings, that is, it’s a measure of how much it would “cost” to transform one line of text (e.g., a word or sentence) into another. The simplest edit distance is Levenshtein distance, where all operations – deleting, adding, and replacing of any character (hence a phoneme in a phonetic transcription) “costs” the same price (usually – one unit). Already in 1995, Kessler compared Irish dialects, realised that such an approach would be too inaccurate and Wagner-Fischer distance should be used instead of Levenshtein distance, meaning that substituting different phonemes should be of different “cost” (e.g., replacing consonant with a consonant would be “cheaper” than replacing with a vowel). However, he confined himself to choosing 12 phonetic parameters, without trying to see how consistent and mutually relevant this twelve-dimensional phoneme space is, as even such an approach gave him sufficiently good results in the categorisation of Irish dialects. We were not satisfied with this approach, because we wanted not only to obtain satisfactory categorisation of Latvian dialects, but also to create a space corresponding to intuitive notions, and with the smallest possible number of axes (Kessler has divided characteristics of a similar nature into separate axes). So we referred to Latvian phonetician Juris Grigorjevs for advice, and as a result, we jointly described such a space.

But first, let's explain what is called “phoneme” in this study. We do not want to change the traditional definition of phoneme. However, if one accepts that a phoneme is merely a functional unit in a particular language, the realisation of which is a sound of a language or a set of them, it comes to a fact that it’s impossible to compare objectively phonemes or phonetic systems of two or more languages. In this case, only phoneme implementations or allophones in the same phonetic neighbourhoods would be comparable in terms of their acoustic, articulate or auditory properties. Such a comparison should be based on the analysis of audio recordings and documentation of the functioning of speech organs using various technical devices.

If you want to compare two or more languages, or several dialects of the same language, you usually try to compare the ideal sound of those languages or dialects that have emerged in the consciousness of speakers and listeners during the communication process. This ideal sound is usually independent of individual or socio-territorially determined pronunciation characteristics and is a kind of a standard for a language user, to which he seeks in his pronunciation. This standard sound is displayed in dictionaries that indicate a preferred pronunciation of each word in a phonetic transcription. Because in practice sounds or their complexes are used for meaning differentiation, parallels can be drawn between a pronunciation representation in dictionaries or other sources in this wider (phonemic) transcription, which distances from a detailed representation of individual pronunciation features of specific speakers in favour of the collective ideal pronunciation, and phonemes of the language or dialect. Although the structure of speech organs of each individual is slightly different, it is clear that, in general, the anatomy of speech organs of all humans (except those who experience pathology of speech organs) is very similar. From this it follows that all people of the world are theoretically capable of uttering all sounds of the world (all languages of the world consists from). It can therefore be assumed that it is theoretically possible to create a universal phonetic system of all phonemes of the world, and users of each particular language just will need to choose certain elements of it to create a subsystem of their language.

[4] The traditional name of Latvians speaking (more or less) official Latvian, mostly from Livonia and Courland.

In terms of set theory, it could be formulated as follows: if we combine phoneme sets from all languages of the world, we get a “supralingual” set of phonemes; and with these “supralingual” phonemes we operate in this study. The International Association of Phonetics (IPA) has created a universal sign system – the International Phonetic Alphabet (also IPA), which can graphically represent phonemes or their specific implementations in any language of the world. The use of a common sign system and the principles of its application make it possible to compare phonemes of different languages or dialects, provided that their systems are sufficiently described in linguistic literature. In the absence of such a theoretical description, it is possible to compare sound systems of idioms by audio material, pronunciation in dictionaries or phonematic transcription. Since our intention was to create a tool that would allow comparison of languages or their subsystems, we named it a “space of phonemes”. By this we mean a universal tool that stands over specific languages and is based in human speech apparatus structure and speech ability. In the current model of space of phonemes, it is not possible to position some of the less commonly used sound groups (clicks, whistles, implicit locks, etc.), as such sounds do not occur in languages with which Latvian is related or connected. If necessary, this deficiency can be overcome by adding appropriate dimensions to the initial space of phonemes. Although this chapter deals with space of phonemes, the method developed is equally applicable to comparing also sounds of various idioms. If an exact description of pronunciation of individuals is used as a basis for comparison, then the pronunciation of different individuals can also be compared in the space of “phonemes” we offer.

To create a space of phonemes, it is necessary to formulate a coordinate system (its dimensions and the zero points of these dimensions) in which Euclidean distances to be calculated correspond to pronunciation and perception facts or the subjective perception of these processes. As acoustic phonetics theory states that the quality of each sound is largely determined by the size of the constriction and the positioning of the resonator, the aperture of articulation and the position of articulation were initially selected as main axes. As early as 2006, Grigorjevs, in his paper “Alternative Classification Model of Latvian Sounds”, used these dimensions to put both vowels and consonants in one, unified system.

The idea of such a system came after reading Ball's article, “Teaching Vowels in Practical Phonetics: The Auditory or Articleulatory Route?”, in which the vowel system model was designed depending on the position of the highest point of the tongue relative to the rim of the palate and the back wall of the throat.

Because vowel quality is determined not by position of the highest point of the tongue but by location and extent of the narrowing of the resonator, Grigorjevs created an alternative vowel classification model. Such a model of the vowel system not only corresponds to their articulatory features, but is also related to the perception-based vowel tonality and bemoliness indices developed by R. Piotrovsky.

A little later, J. Grigorjevs created a table in which both vowels and consonants can be arranged logically by size and position in resonator of the articular aperture. Initial dimensions of the space of phonemes were constructed on the basis of the principles presented in this table. Value of “occlusive” was chosen as the zero point of the axis of articulated aperture, because at the event of full closure the airflow path is blocked and the aperture equals 0. As the aperture increased, values on this axis were defined by a specific quantitative step and named “constrictive fricative”, “more open constrictive”, “narrow vowel” and “more open vowel”. However, when phonemes were arranged this way, in many cases, distances between phonemes traditionally assumed to be close were greater than those assumed to be distant (e.g., the phoneme /n/ was closer to /l/ rather than /ņ/), and distances, that intuitively seemed to be equal or very similar, differed significantly (e.g., the distance between /n/ and /ņ/ was one and a half times bigger than between /l/ and /ļ/). Such a non-compliance pointed to the fact that the two dimensions used characterizes phonemes’ differences insufficiently, and apparently the number of dimensions needs to be increased.

After careful research, we realized that three-dimensional space had to be constructed, dividing the former axis of articulation into two: the “non tongue-backside” axis (neutral-alveolar-dental-labial) and the “tongue-backside” or “softness” axis (pharyngeal-uvular-velar-palatalised-palatal). This is due to the fact that consonants that are not pronounced with the backside of tongue may have additional articulation with the backside of tongue, giving them a “lighter” or “darker” sound. So, the back of tongue is used to pronounce vowels and to pronounce some consonants – both to create a gap or a closure, and to modify the shape of resonator shells in a case of other articulations. In addition, account should be taken of the fact that there are consonants with double articulation in languages of the world, which usually involve one articulation point on the back of the tongue and another point on the tip of the tongue or lips, and therefore an opportunity of description of such consonants in the system being developed should be provided. After this modification our phonemes’ space model became closer to the APEX speech articulation model developed at Stockholm University and the Royal Institute of Technology (Stockholm), which is based on long and careful research on sound pronunciation and acoustic parameters. In the APEX model, two of the main parameters for determining sound articulation are the parameters of the body of tongue and the tip of tongue.

Authors of APEX have chosen the initial position of the body of the tongue that corresponds to the velar articulation, from which the body of tongue can be moved forward to the palatal articulation or back and down to the pharyngeal articulation. Because the tongue’s body parameters in the APEX model determine the articulation site of the backside of tongue at the passive speech organ, we chose the velar tongue backside as the initial position in the “tongue-backside” or “softness” dimension, denoting it a neutral value of 0. Moving the back of tongue towards the hard palate, gradually increases the “brightness” and softness of hearing, so the positive numerical value increases from “velar” (0) to “palatalised” (1) and “palatal” (2). Moving the back of tongue towards the throat or the lower part of the pharynx makes the sound “darker”, hollow, harder, so the negative value in this direction increases from “velar” (0) to “uvular” (-1) and “pharyngeal” (-2).

In the “non tongue-backside” dimension we combined positions of the articulation defined by the tongue tip’s and lips’ parameters in the APEX model. As a starting point, we assumed a free tip of tongue when the tip of tongue is not actively involved in the wording of sound, calling this articulating position “neutral” and assigning it a value of 0. As the articulation position moves forward from the maximal anterior (palatal) back of tongue articulation, the value of this dimension is gradually increasing in the direction from “neutral” (0) to “alveolar” (1), “dental” (2) and “labial” (3). In such a system, should be assigned values of 1.25, 1.5, 2 and 2.5 to denote laminal or apical dental, interdental and labiodental sounds, respectively. If retroflex consonants were to be included as separate phonemes in our phoneme space, the “tongue-non-backside” dimension would be supplemented by a value of -1, since the articulation of these consonants occurs with the bottom edge of the tip of tongue, folded up and back.

In addition, to create a space that fully describes all phonemes of Latvian dialects, that is, any different phonemes have different coordinates, five more dimensions were created: labiality (neutral-labialized), trembling (neutral-vibrant), non-vocality (vocal-non-vocal), transversality (medial-lateral), and nasality (neutral-nasal), thereby obtaining an eight-dimensional space. The labiality dimension makes it possible to distinguish between sounds that are not pronounced and those that are pronounced with lips’ stretching or rounding. Because in most cases lips plays a passive role in sound pronunciation, this pronunciation was classified “neutral”, giving it a value of 0, but a “labialized” pronunciation with actively rounded or stretched lips, a value of 1. The dimension of trembling was introduced to distinguish (otherwise similar) vibrants. If a consonant is pronounced without vibration of a speech organ, its articulation is classified as “neutral” and assigned a value of 0, but if a consonant requires vibrations of a speech organ, it is classified as “vibrant” with a value of 1. Vocality of sound is characterized by “non-vocality” dimension, in which vocal sounds were recognised as neutral due to the majority, assigning them a value of 0, but for non-vocal sounds – a value of 2, leaving some place for intermediate values. The transversality dimension is required for separation of lateral consonants. If during a consonant articulation, the air flow moves in the longitudinal direction of a resonator formed by speech organs, the articulation is classified as “neutral”, assigning it a value of 0. If an obstacle is created on the path of the air flow, by bypassing which it is directed sideways, the articulation is classified as “lateral” and gets a value of 1.

The nasality dimension makes it possible to distinguish oral sounds from nasal sounds. If sound articulation occurs through the mouth, blocking the airflow path through the nasal cavity, as is the case with most sounds’ articulation, the articulation is classified as “neutral” by assigning it a value of 0. If the air flow path through the nasal cavity is open during the phormation of the sound, the articulation is classified as nasal, assigning it a value of 1. The classification of nasal consonants provoked a reflection on its inconsistency with the traditional one, when they are called occlusives. Associatively, an “occlusive” is associated with a complete interruption of airflow, even for a short period. Of course, during a pronunciation of nasal consonants, a passage of air through the mouth is blocked, but the flow of air and sound passes relatively freely through the nasal cavity, producing additional and characteristic resonances of nasal sounds. This free airflow is the reason to classify nasal consonants as sonants. In view of these characteristics, we classified nasals as “more open constrictives” instead of “occlusives”.

Based on the phonetic properties of the phonemes found in Latvian dialects, we were assigned relative coordinates to each of them which, despite their conditionality, represent the closeness of the phonemes well enough, respectively, corresponds to traditional, intuitive notions. In addition to phonemes, such coordinates were also assigned to variants of phonemes or allophones found in dialects and subdialects.

To be able to calculate a value not only for replacing sound, but also for deleting or adding it, it is also necessary to define a so called neutral point, or ε, in this case it could be called coordinates of the blank sound too. We considered three possible solutions: the origin of the coordinate system (corresponding to the consonant /g/), the midpoint of the intervals used (such a combination of parameters would correspond to a mysterious, unspoken sound), and a physiologically based, specially defined point. The third way seemed the most reasonable, so we decided to choose it.

However, the selection of this physiologically based point selection was not unambiguous too: for example, coordinates of a point corresponding to steady breathing through a nose would be (2, 0, 3, 0, 0, 0, 0, 1), but for steady breathing through a mouth – (2, 0, 0, 0, 4, 0, 0, 0). Comparing the coordinates of these points with the coordinates of the phonemes shows that the first is closer to the consonants than the vowels, and the second is vice versa. This can be quantified by calculating distances. Since, in our intuitive perception, vowels are pronounced with less effort, while consonants are difficult to pronounce without a vowel sound, it seemed more reasonable to choose a zero point closer to the vowels. Therefore, we chose the breathing through the mouth: ε = (2, 0, 0, 0, 4, 0, 0, 0).

In the PERL programming language, we developed a program that calculates the Euclidean distance between all pairs of phonemes according to the coordinates defined above and displays the results in tabular form in a web browser window. The program allows you to interactively change weight factors of axes, respectively, to reduce or increase the effect of the dimensions on the resulting distance. As experimenting, we have come to a conclusion that a table with the “softness” dimension weight factor of 0.5 corresponds better to an intuitive understanding of distances between the phonemes, leaving other weight factors at the default value of 1.

Method of Edit Distances for Parallel Sets of Words (Vocabularies)

We have discussed above the n-gram method for arbitrarily selected texts. One of directions in which to advance would be to move from arbitrary texts (corpora) to parallel texts (corpora), thus reducing the amount of data required to obtain positive results. Of course, collecting such parallel information on expeditions would still require a large investment of time and work, but it turned out that for Latvian subdialects were parallel sets of words already collected, admittedly they are kept in notebooks and therefore need to be digitized. In 1954, Institute of Language and Literature of Academy of Sciences of Latvian Socialistic Soviet Republic published a book “Program of Collection of Materials for Dialectology Atlas of Latvian Language”, where were described not only a specification of the phonetic transcription system, but also a list of questions to be answered for each subdialect speakers.

This list contains 670 questions, 103 of which have been drawn up for showing phonetic, 160 – morphological, 107 – syntactic and 300 – lexical differences. As in explanations of the book is written: “According to the mission of the dialectological atlas – to give a synthetic overview of linguistic phenomena that differentiate between dialects as well as subdialects, the programme attempts to include phenomena of Latvian phonetics, morphology, syntax and vocabulary that are spoken in different regions.” Since we are also primarily interested in differences of idioms rather than matches, such materials are fully relevant to our needs and can be used in our research.

The next step was to choose a suitable comparison method. It turned out that Irish and Dutch colleagues had already worked on similar data, so we decided to try recommended by them method of defining a distance between languages as a sum or an average of Levenshtein’s distance between words of the same meaning (in their phonetic transcription) and then using the hierarchical agglomerative clustering of average distance for analysis of the results obtained.

Levenshtein distance is called the cost of re-editing one line of text into another, where operations or so called edits are insertions, deletions, or substitutions of a character, but cost of each operation is defined equal to 1. That is, the price equals the minimum number of editing actions. For example, to get the word “doktors” from “maģistrs”, we have to: delete “m” and “a”, replace “ģ” with “d”, “i” with “o”, “s” with “k” and insert “o” at the appropriate location. Thus, the Levenshtein distance between these words is equal to 6.

We inputted into computer a small sample of data – 13 words, representing phonetic, 10 – morphological and 8 – lexical differences, all for 13 subdialects from different parts of Latvia. In order to facilitate input (which should be possible from any console) and processing (it is important to code all phonemes with a same fixed number of bytes), the data was transcribed in a phonetic pseudo code where each phoneme was transcribed with 6 bytes (1, 2 – the phoneme itself, 3 – length, 4 – intonation, 5 – stressness, 6 – syllabicity).

In the PERL programming language we wrote a program that calculates average Levenshtein distances between pairs of subdialects (i.e., averages between given pairs of word entries of these subdialects) and then clusterises the results.

After several trials we found out that the hierarchical agglomerative clustering of average distance sometimes works inappropriately (in cases where proximity of two subdialects is obvious, and one of them is in a group, but the other one – outside the group, by the arithmetic mean the second one is calculated as a closer one), and therefore, according to the characteristics of the objects under consideration, it would be better to use the hierarchical agglomerative clustering of smallest distance, that is, when two categories are merged into one, we define the distance between the new, merged category and any of other categories as the smallest distance, rather than the arithmetic mean, between each member of the new category and the other category.

The results of the experiment proved to be good enough and were consistent with the notion of the closeness of the selected dialects. However, we decided to make our experiment more sophisticated by moving from the Levenshtein distance to the Wagner-Fischer distance, namely by introducing a price that depends on the proximity of phonemes instead of a constant phoneme replacement price. (For example, a price of /ɣ/ substitution for /g/ obviously should be less than substitution for /b/, but it should be less than price of substitution for /a/.)

This is how we came to the problem, which we have already addressed in the previous chapter – the creation of the phoneme space. During this experiment, though, the development of the space of phonemes was only in process, so we did it with its “working version” – not the eight-dimensional, but a six-dimensional (vocality, softness, tongue-non-backside, labiality, aperture, and trembling).

We had also chosen the zero point purely mathematically – as the midpoint of the axes we use, not the phonologically based one we figured out later.

We also expanded each phoneme’s description with number of parameters that characterized not so much the phoneme itself as its applications: length of the sound, intonation (we defined a binary distance because there are five intonations in the idioms we analysed, but it is not clear which are further from each other, and to what extent), stressness, and syllabicity.

We created a script for calculating Wagner-Fischer's distance on the basis of Levenshtein distance’s script.

The use of Wagner-Fischer distance didn't bring significant changes compared with Levenshtein distance. However, there is such a-such nuances, that intuitively seems reasonable are. Thus, for example, with Wagner-Fischer distance, two Selonian subdialects of different regions (Leivuons – Latgalia, and Rite – Courland) formed a pair by both phonetic and morphological questions, but lexically Leivuons subdialect appears to be a little be closer to other Latgalian subdialects (which looks quite logical, because there is a lexical layer like the Catholic one, which is specific to Latgalia).

At first glance, the morphological closeness of Atašine and Zīmers, which appears only with Wagner-Fischer distance, may seem unexpected, since Atašine is in Latgalia and the subdialect there is a deep Selonian one, while Zīmers is in Livonia and the subdialect there is a deep Latgalian one, besides these places are very far from each other geographically too. However, if we think about it, we can find an explanation: both subdialects have been under a great influence of Middle Latvian subdialects for a long time, and it is the morphology that has been most influenced by it.

There are also such vivid examples of separation of question groups that occur both with Wagner-Fischer and Levenshtein distances, for example, in Ģeri people speaks in Livonia’s Livonian subdialect, so it is not surprising that it appears phonetically closer to Courland’s Livonian subdialects, but lexically – to their neighbors, Kocēni Livonia’s middle dialect’s subdialect.

Of course, the very fact that the results of the categorisation of morphological and lexical questions coincide indicates less nuance in the results obtained with the Levenshtein distance. We can therefore conclude that it is worthwhile to use Wagner-Fischer distance for this type of measurement when we are interested in linguistic nuances, but using the Levenshtein distance justifies when coarse but rapid categorisation is needed.

We published this method in 2006. Many years later, in 2014, colleagues at the University of Helsinki published an article of a similar direction. It uses data from the multilingual etymological database *Starling*, created by the outstanding Russian linguist Sergey Starostin. In this case the data is actually retrieved in a form of phonetic transcription of parallel records of multilingual dictionaries, thus the same as in our experiment. Normalised compression distance is used for evaluation, which is a new and interesting solution and should be tried in the future with our data too.

Method of Phoneme Recognisers for Full-size Recordings

So far, we have been working separately – either with phonetic transcription or straightforward with speech recordings. In this chapter, let’s look at a method that links these types of data, which are so related by their nature, in practice too.

The emergence and development of phoneme recognisers for spontaneous speech recordings led us to the idea of a combined method: first, phoneme recognisers generate appropriate transcriptions from speech recordings, and then methods that have already been used for manual phonetic transcriptions are applied.

At first sight, it would appear the setting of the method is inherently wrong, as it is not yet known about an existence of a universal (i.e., a language-independent) phoneme recogniser – all the recognisers we know have been trained (i.e., models created for them) on specific languages and, with strict veracity, are also designed for these languages. But in talks with colleagues involved in the development of the recogniser *PhnRec* that they experimentally tried to identify phonemes of other, non-model languages too, and the results have proved to be good enough, albeit worse, than for a model language.

It was also confirmed by the results published by the developers regarding use of phoneme recognisers in language identification, because, as in our case, it is also a solution that uses phonotactical methodology (although they use data from all potentially identifiable languages for training). However, the possibility of using a monolingual phoneme recogniser is also described and interpreted: it simulates a situation where a monolingual person hears different foreign languages.

Therefore, we allowed ourselves to hypothesize that using a phoneme recogniser modeled on a language for a pair of other languages might not significantly affect the phonotactic characteristics we use to determine a degree of proximity.

We had at our disposal spontaneous speech recordings of five Latvian subdialects. All recordings were collected in accordance with high principles of gathering, it means, all records were uniformed, recorded with the same type of hardware (a dynamic one-way microphone fixed on heads of speakers was used), an external noise was minimized as far as possible. All entries were manually cleared – i.e., all other voices and sounds were cut out, leaving only the speech of the main speaker. Recording technical quality was 44.1 kHz / 16 bit. Depending on the requirements of the particular model of phoneme recogniser (in this case, the technical quality of the training recordings), the recordings were downsampled to either 8 kHz or 16 kHz.

PhnRec is a phoneme recogniser developed by the Speech Processing Group of the Faculty of Information Technology, Brno University of Technology. Its main feature is the use of long time context (up to several hundred milliseconds). The retrieval of speech characteristics is based on the division of time context, the role of the classifier is performed by artificial neural networks, and the decoding of queues is performed by the Viterbi algorithm. The program package includes already trained Czech, English, Russian and Hungarian language models. For the sake of greater comparability, we decided to try them all.

Software package *Sphinx 4* includes a program *pocketsphinx*, which works as a phoneme recogniser too. Unfortunately, we were unable to find a publication describing the phoneme recognition algorithm used by *Sphinx 4*, and we did not have time to investigate the program code ourselves. From the overall documentation, it was possible to understand that the program creates a hidden Markov model for each phoneme during a language model training (i.e., a time context taken into account is not long), but decoding is done not only by the “classical” Viterbi algorithm but also by the Bush-Derby algorithm.

By default, the package contains only an English language model. Other language models from third-party developers are available, however we decided to use only models developed by the recogniser developers to exclude errors that might result from a poor quality of model development.

Earlier, we described the n-gram method and its application for phonetic transcriptions designed by manually transcribing sound recordings. As the results were good, we can also apply this method to transcriptions created in an automated way. Let us recall that we have proved that the distance defined within the n-gram method is a metric, so in fact, the distance obtained by this combination method will be a metric too.

The experiment was done with scripts written in PERL programming language. Initially, we applied all the recognisers and models at our disposal (i.e., 5 in total: *PhnRec* for Czech, Russian, English and Hungarian, as well as *Sphinx* for English) to the recordings available to us. Then we converted the obtained phonemes files into a two-byte phonotext, and created a file for each subdialect, merging together all the informants (i.e., speakers) of this subdialect. And then, with the help of modified and corrected by us PERL-program *TextCat,* we calculated distances between these newly created phonetic texts. The results were quite good: subdialects, which are closer by intuitive understanding, were closer by our distances too.

For greater visibility, we decided to categorise our results using the method of hierarchical agglomerative clustering of smallest distance. As a result, we obtained two types of trees. Both separated North Latgalian subdialects in a separate branch (which was really expected given their great similarity). The only dialect of Courland was also singled out (it certainly could not have been otherwise). The difference appeared with middle Latgalian dialects: the first tree (according to the results of Russian and Czech models) putted them into a common branch, while the second – not. From an intuitive and linguoanalytical point of view, the first variant seems to be more correct, though the second one has some meaning too. Since Russian and Czech are Slavic languages, but according to the genealogical classification of languages, the closest for Baltic languages are Slavic languages, we can conclude, that results of the method of phoneme recognisers are better if a model's language is more closely related to a recordings' language.

First of all, we proved that our basic hypothesis is confirming: phoneme recognisers are applicable to speech corpora for automated language proximity estimation.

Secondly, we made sure that the training language of a model influences results, but not to the extent that the application is excluded. That is, in principle we can use any recogniser with a model for any language, but for the purity of the experiment, it is desirable that the degree of affinity between a model's training language and various languages of the experiment is not too different.

Thirdly, phoneme recogniser's architecture does not significantly affect results. (Although the recognition results of *PhnRec* were more "readable".)

And fourthly, we have actually defined a metric in the languages (idioms) space, and moreover, it can be calculated automatically, without any significant data preparation, i.e. without manual or so-called "black" work. It looks very promising.

Method of Hidden Markov Models for Full-size Recordings

We chose the undivided full-length speaker’s sound recording as the object of the HMM and built models on a sufficient number of informants’ (i.e., speakers’) records of a language (idiom).

It is clear that such a method is applicable to any human language, including ones without a written form. However, since Latvian – both Baltian and Latgalian – subdialects were more accessible to us, we decided to experiment with them.

In autumn 2008 we went on several expeditions in Latgalia and Courland. As a result, 4 Latgalian and 1 Couronian subdialects’ material was collected: 30 informants who spoke in subdialect of Vileks, 23 of Baļtinova, 29 of Rudzātys, 14 of Auleja and 17 of Dundag. All the informants told their life story: about parents, grandparents, brothers, sisters, family, school, work, wedding, children, farm, etc. Part of the material collected material we used in our experiment.

In fact, several experiments were carried out to find out and test the proposed method. They were all implemented by the help of HTK package, i.e. there was no need to program the algorithms and even to study their implementation in the package, since it is recognized among speech researchers worldwide. Of course, some scripts were developed for data processing and automation purposes.

The first experiment was carried out with a read speech: the same text read by the same person in three languages – Latvian, Latgalian and Russian. Four recordings were recorded in each language: three were read at medium speed and one – at accelerated; length of each of the recordings – 1 to 2 minutes. For each language on all the three medium speed’s speech recordings hidden Markov model was created. After that with the HVite utility (the implementation of the Viterbi algorithm in the HTK package) the nearest model for each of the high-speed speech recordings was founded. With a small number of Gaussian mix components (so-called “mods”) the results were unsatisfactory, but with four and above worked properly – the high-speed speech recordings’ languages were detected flawlessly.

The positive results of this experiment motivated us to do the next one, this time on the real data of our research.

We chose two from our recorded dialects – Rudzātys and Vileks, both Latgalian, but from opposite sides of Latgalia: Northeast and Southwest. Thus, the chosen languages were very close (and it, of course, reinforces the importance of results in a case of a positive outcome), but at the same time far enough to be sure that differences will not be smothered by social contacts of speakers. From each language we randomly chose eight female informants, those eight were randomly divided into two subgroups: five for model creation and three for testing. The results were identical to the results of the previous experiment: in the case of a small number of mods, languages were detected erroneously (in different ways, without understandable consequences), but in case of four or more – flawlessly.

Thus, we can conclude that our hypothesis of the possibility of training HMMs on full-size recordings to describe language as such was confirmed. We assumed that once it works in recognition tasks, i.e., the language of other recordings is correctly determined by such models, it should also work in determination of the distance be-tween languages, i.e. we could define a distance between languages as a distance between HMMs of these languages.

That's why we decided to create HMMs for all the five of our dialects and define different types of metrics in their space.

Initially, we decided to try our luck with the well-known metric – Euclidean. Then, the choice was made as for the data (characterising the distribution) that would be dimensions of our metric space. It seemed reasonable to use mean value vectors (model includes mean, variance and weight vectors).

Firstly we made distance calculations for the above mentioned Latvian/Latgalian/ Russian read speech. We calculated Euclidean metrics, normalized Euclidean metrics (normalized by the first, second, and both arguments) and Jordan metrics.

In case of correct distances one should expect that distances between speech samples of the same language are smaller, between Latvian and Latgalian – medium, between Russian and Latgalian – bigger, and between Russian and Latvian – biggest ones. However, for all the three metrics, it can be seen that the distances are very similar, and at the same time they are "jumping" – having unpredictable changes, that makes possible, that intuitively closer languages have larger distances and vice versa.

We carried out this experiment on our dialects' speech samples too.

Unfortunately, the program HERest from the HTK package, which performs a recalculation of HMM parameters using the Baum-Welch algorithm, obviously has a fault – at a larger number of input files, it displays an error message that approximation cannot be calculated: *WARNING [-7324] StepBack: File [path] - bad data or over pruning*. Such a problem should occur if the recording is technically poor or has some other fault. However, it is interesting that for a same file this error could appear with a larger number of files, but not appear with a smaller one – hence it does not depend on the file quality, but on something else. This leads to the conclusion that this is a fault of the program, and the only way to avoid it is to bypass it. As we simply did not want to skip some of the files, we decided to divide the voices of men and women into separate groups – there were fewer files in each group and HERest stopped crashing. Thus, the experiment became larger and probably more interesting, but it also has one drawback – we will not be able to compare directly its results with results of other methods.

As we can see, all the distances here are "dancing" – "men" of the same language sometimes are further than "women" of other language, intuitively close languages sometimes appear further than distant ones.

At the suggestion of Professor, *Dr. habil. math.* Aivars Lorencs, we decided to try the same metrics, but for the mean values divided by the variances, that is, the more volatile are values, the smaller is a weight – they are affecting less a value of the distance.

As we can see, in any case, namely, for any data set and any metric, this improve-ment has not made results consistent.

That's why our conclusion is negative: we cannot define the distance in this way and should look for other ways to do it.

The most common assessment of HMM similarity is the Kullback-Leibler divergence, which the authors have been defined in their publication of 1951.

It is a mathematical expectation of a logarithmic difference between two probabili-ties distributions by the first distribution. So, naturally, it is not symmetrical, so it does not correspond to one of the axioms of metrics and is not a metric. Defining an arithmetic mean of divergence values of both directions often solves this problem.

Kullback-Leibler divergence was calculated with a slightly modified Python script written by Speech Lab of Technical University of Brno.

At first glance, we can see a certain coherence in the results (e.g., the fact that Dundag looks further, or the fact that Baļtinova and Vileks is the closest pair), though, of course, the lack of symmetry and the separation of the voices of men and women is confusing and does not allow to analyze the results properly. Therefore, we decided to simplify them: first, to symmetrize the table by calculation of average arithmetic values and, secondly, to put together the male and female voices, also by taking the average arithmetic value.

This has brought all the values closer, which confirms that such a great range of values had other reasons than the qualities of languages. This, of course, is not good. However such similar values might reflect something – so let's look at them.

The distances of Auleja looks adequately: Dundag – the furthest, Rudzātys – the closest, Baļtinova closer than Vileks.

The results of Baļtinova could also be considered (Vileks very close, Rudzātys further) well if it were not for the unjustified Dundag's proximity to Auleja.

Even worse results for Rudzātys – Baļtinova appeared to be closer to Auleja, Dundag –

closer to Vileks.

In contrast, Vileks looks very good – Baļtinova is the closest, then Rudzātys, then Auleja, and Dundag the furthest.

In general the method is usable. However, it is technically complex and the results are not fully reliable. Therefore, other methods are more recommended for real use.

Method of i-Vectors for Full-size Recordings

i-vectors are a relatively new way of solving identification and recognition tasks, which are now being used to identify other types of objects too, but they were originally thought of in search for a speech recognition tasks.

The first widely known publication in which this new idea was expressed (in terms of speaker identification) was in 2009, though the name of “i-vectors" has not yet appeared, but their space was named “total variability space”. In early 2010, the name “i-vectors” appears as an additional one, but by the second half of that year, it is already in full swing – already describing a challenge of language identification.

The i-vector method is based on the representation of Gaussian mixture models of expressions with a hidden low-dimensional variable and the use of this expression representation as a feature vector in the language classifier.

There can be different i-vectors depending on what linguistic information they contain, for example, acoustic, prosodic, phonotactic, on what kind of data they are built on – continuous or discrete, and for what they are intended for – speaker identification (SID), language identification (LID), or other tasks. So, in fact, we could even talk about a whole bunch of methods, but immersing in such subtleties is not a goal of our dissertation.

In our experiment, we used the set of dialect speech recordings described in the previous chapter.

Since we had scripts for i-vector calculations developed by Brno University of Technology we used them, of course, instead of developing them from scratch. In 2015, BUT Speech Group came up with a proposal to create a common Voice Biometry Standard, or VBS, as various technical standards currently in use do not allow rapid provision and exchange of data. The standard package also includes Python scripts for i-vectors’ computation (they only require voice recordings to operate, but better results can be achieved if there are already externally defined voice activity intervals, so-called, VAD, or “Voice Activity Detection”), because the built-in activity detector is very primitive.

Sure, the biometry standard is designed for speaker identification tasks, that is, it focuses on specific characteristics of a speaker, i.e. a particular human individual speech (including individual voice particularities), and so i-vectors generated by this package are called SID (or “Speaker IDentification”). They are theoretically less suitable for our task, but we decided to give them a try because open standards and public availability of scripts are key factors in technology choice.

As with HMM in the previous chapter, i-vectors were calculated for full informants speech recordings, thus expecting them to be representative of the language as a whole. For all pairs of dialects, we calculated the cosine similarity between the resulting i-vectors.

The results were quite good. It may need to be clarified here that the cosine similarity is a cosine value, and an arccosine gives an angle. Angles, in turn, are easier to visualize: imagine a zero line drawn in a plane and the point or centre of it; then an angle formed by rays from this center and the zero ray (the right ray of the zero line) represents corresponding distances between subdialects – the smaller the angle, the closer the tongue is.

Thus, Baļtinova and Vileks appear to be the closest pair of subdialects (47°). Also, the distance between the South and West Latgalian subdialects – Rudzātys and Auleja – is smaller than between them and North Latgalian subdialects. The only assessment that seems to be very wrong is the distance between Rudzātys and Dundag subdialects: it certainly did not have to be smaller, and so much smaller, than the distance between Rudzātys and the other three Latgalian subdialects.

We decided to try Euclidean and other vector metrics (e.g., Jordan) for i-vectors. We also decided to try city block metrics. To apply such metrics, we calculated arithmetic mean i-vectors for each language from its informants’ i-vectors.

The worst of these metrics (though not really bad) in our case turned out to be Jordan: it shows both Vileks and Rudzātys much further from Auleja than from Dundag.

The L_1 and Euclidean (both non-normalised and normalised, since normalisation does not change the substance of the case) metrics look equally good and better than cosine similarity: Vileks and Baļtinova are the closest, Dundag to all Latgalian dialects – the furthest. The only question that arises is: why does Auleja to Baļtinova appear to be closer than to Rudzātys? As if it shouldn't be. This could be a fault of metrics, a deficiency of the data, but also an objective estimation that takes into account some of nuances of subdialects that are usually overlooked in theoretical comparison.

During the internship at Brno University of Technology we gave to Speech Lab’s researcher Oldrich Plot the speech corpus we collected, because he asked them for his experiments, and after a while we got results.

Prior to conducting the experiment, the data from each subdialect were randomly divided into two parts: a larger training part and a smaller test part. Then, i-vectors were calculated using each part separately. Then, the Gaussian linear classifier was trained on the training part of i-vectors, and on the test part of i-vectors it was applied. In terms of percentage distributions of how much of the test data was correctly distributed and how much of it gaved to other dialects, the results were quite good: Dundag as the most distinctive is recognised bestly; Rudzātys is also very good, and the fact that it “gives” a part to other Latgalian subdialects is just natural; Baļtinova and Vileks, in view of their closeness, also show relatively good results, with most of the difference being “returned” to each other; the only thing that surprises is Auleja's comparatively poor performance “in favour” of Rudzātys.

Given the good results of the Brinneners, we decided to try our SID i-vectors experiments on their LID i-vectors. At our request, they kindly gave us these i-vectors. We expected the results to be similar to the SID i-vectors, but slightly better, because LID i-vectors are designed to perform a more similar task.

To our great astonishment, the cosine similarity was completely meaningless. We are still working on determining the causes.

However, we decided to try the other distances used for the SID i-vectors as well. The Euclidean and L_1 metrics were similar to those of the SID. Interestingly, however, that Jordan metrics, which was not so good for SID i-vectors, behaved much better for LID i-vectors – did not had such problems as SID, and could be said to be almost equivalent to Euclidean and L_1 for LID.

We are convinced that i-vectors are good enough to represent languages and can therefore be used to quantify language differences. In addition, they are more convenient to use than hidden Markov models: both in terms of data convenience and popularity of metrics could be applied. According to the results of our experiments, we recommend using Euclidean or L_1 metrics.

Method of Gaussian Mixture Models for Full-size Recordings

Another type of speech modeling is a use of Gaussian mixture models (GMM).

A Gaussian mixture is a sum of a finite number of Gaussian or normal distributions, more precisely – a weighted sum (with scalar weights). Such a model is described by three variables – the mathematical expectation vector, the covariance matrix and the weighting factor vector. Since a sum of independent normal distributions is a normal distribution, the Gaussian mixture is a normal distribution too.

GMMs are widely used to classify data (e.g., in economics, demographics, ecology, etc.) when there is a reason to believe that each of these classes corresponds to a normal distribution. Thus, GMMs are also used in different types of recognition tasks (image, sound, and other types of objects) because by their very nature they are the most probable belonging to a given class.

Since we are interested in speech modeling, we worked with distributions of spectral information features of speech signals. The feature vectors are constructed from the Fourier cosine transform coefficients, i.e. MFCC (*Mel Frequency Cepstral Coefficients*), using 14 coefficients per frame.

Models were created using the expectation-maximisation algorithm or EM, but in two ways – as fully independent, individually trained models, each built on speech data of only one particular subdialect, as well as on a common model, created on data of all the subdialects, so-called universal background model or UBM with MAP adaptation.

In the experiments of this chapter were also used the set of dialectal speech recordings described above.

And in this case too, we used the technique already used by other methods – to train models on full informants' phonograms. We did it both for full data (which varies in size for different subdialects) and for data of the same amount: for subdialects with a bigger amount of data, all recordings were truncated proportionally (by cutting their end parts).

Experiments were performed using capabilities of *MatLab* package. The scripts were developed by us on the basis of information provided by Jose Benito Trangol, a PhD student at the Speech Laboratory of the National Autonomous University of Mexico, on his experiments in the field of speech recognition and libraries, functions and parameters used.

As the **first** experiment, we calculated Gauss mixture models directly, without adaptation, i.e. each model was composed solely of recordings of the corresponding subdialect and completely independent of other models.

In this and subsequent experiments, we calculated Euclidean, L_1 (or city block), Jordan (or Chebyshov) metrics, and Kullback-Leibler divergence for all the three model components: mean vectors, covariance matrices, and weight vectors.

All in all, the Euclidean metrics' values correspond to an intuitive notion of the proximity of these subdialects: Dundag is the furthest from all other ones, Vileks and Baļtinova are closest to each other, and so on. It is somewhat strange that Rudzātys and Auleja are further apart than both Rudzātys and Auleja with Vileks and Baļtinova. However, there may be some explanation for this, since the Auleja subdialect is quite specific, whereas both Rudzātys and North Latgalian subdialects contain some similar forms (for example, *tuo* instead of *tai*, etc.).

In contrast, the Kullback-Leibler divergence values look completely meaningless because they contain negative values. City blocks metrics in this case behaves the same way as Euclidean metrics. Jordan metrics looks very shabby: Dundag is not the furthest of all four Latgalian dialects, and Vileks and Baļtinova are not the closest.

Now let's move from the mean value vectors to the **covariance matrices**. The Euclidean metrics behaves just like in case of the mean vectors: all is well, only minor issues are raised by Rudzātys and Auleja relations on the background of North Latgalian dialects.

The Kullback-Leibler divergence looks more or less adequately, but it has minor inconsistencies, for example, for Auleja Vileks is further than Dundaga. As there is no such problem in the other direction, the possibility of symmetrying divergence immediately comes to mind. However, the symmetrying only reduces this problem, not completely solves it.

The city blocks metrics looks relatively good, but slightly worse than the Euclidean metrics, for example, the distance Auleja-Vileks is closer to Auleja-Dundag than to Auleja-Baļtinova. To our surprise, covariance matrices also turn out to be good the Jordan metrics, even though it was completely meaningless for mean vectors.

The Euclidean metric for **weights vectors** also gives a more or less consistent impression, but it also contains significant errors: Auleja to Dundag is closer than to Vileks and Baļtinova; Vileks to Rudzātys is closer than to Baļtinova. The Kullback-Leibler divergence shows the same disadvantages as the Euclidean metrics, and it is clear that symmetry cannot help because they are both ways.

The city blocks metrics also shows the same discrepancies as we saw in the case of the Euclidean metrics and the Kullback-Leibler divergence. The results of Jordan metrics differ: it does not show the first problem, but the second, unfortunately, shows a stronger expression: for Vileks Baļtinova appears further no only than Rudzātys, but even for all other subdialects.

The distance of Mahalanobis shows no signs of any meaning: for Auleja Vileks does not have to be the furthest, besides so very much – several times, for Baļtinova Dundag does not have to be the closest, and also so very much – the difference between the distances Dundag-Baļtinova and Dundag-Rudzātys is almost 200(!) times, etc. Besides, it is not symmetrical, so it’s not a distance at all. Also, the so-called special Euclidean distance is completely pointless: although the proportions are not so blatant, the results are essentially similar to the Mahalanobis case.

As a **second** experiment, we decided to do all the same calculations, but with MAP adaptation. To this end, we created a common background model that was built on 30% of all speakers recordings of all subdialects. Then, with the remaining 70% of the speech data of each subdialect, this background model was adapted and models of the subdialects were created. In other words, during the data preparation process, all files were divided into two parts: 30% long starting parts and 70% long final parts.

In theory, the adaptation can be done not only by all the three, but also two or one of GMM parameters, most often – by mean value vectors. However, in this case, results are not interesting – distances based on differences of mean vectors coincide with distances from models adjusted by all the three parameters, while distances based on differences of covariance or weight are zero. Therefore, we chose to experiment with models that are adapted by all the three parameters.

Interestingly, the Euclidean distance for the **mean value vectors** with MAP-adapted models is worse than for the directly built models: Dundag is “closer” than Auleja for Baļtinova and Vileks. With the Kullback-Leibler divergence something is completely wrong – it returns negative values.

City block metrics in this case face the same problems as Euclidean metrics. As usual, Jordan metrics behave differently and this time – also better: Dundag is the furthest of all Latgalian subdialects. However, there is a small problem: for Vileks Rudzātys appears closer than Baļtinova.

Euclidean distance between **covariance matrices** has the same problems as it had between mean value vectors, but this time they are seen in a stronger form. Kullback-Leibler divergence also has the same fault this time. And the values are almost symmetrical, so symmetrisation won’t save this time.

City block metrics this time have not only all the same flaws as Euclidean metric, but even one more: for Vileks Rudzātys appear closer than Baļtinova. Jordan metrics has the same fault as Euclidean and other metrics, however it appears with other sets of subdialects: Dundag appears closer to both Auleja than Vileks and Vileks than Auleja, whereas for Rudzātys Auleja and Vileks are further than Dundag.

Now let’s move to the **weight vectors**. Euclidean metrics show unexpectedly good results: Dundag is furthest from all other subdialects, but the distance between Vileks and Baļtinova is the shortest. Interestingly, contrary to what was observed in the first experiment, Euclidean metrics shows better results for the weight vectors than for the convergence matrices. In addition, its results are more similar to those of covariance matrices without adaptation. We have no logical explanation for this, since during MAP adaptation process weight vectors does not accumulate data of covariance matrixes of adaptation data, which could produce this effect.

Kullback-Leibler divergence has minor problems: for Baļtinova and Vileks Rudzātys appear a little further than Dundag. Because everything is right in the other direction, let’s try to correct it with symmetrisation. It really helps this time, and after that, Kullback-Leibler divergence looks very good.

City blocks metrics behave similarly to Euclidean, but a little worse: Baļtinova turns out to be a little closer to Dundag than Rudzātys. Jordan metrics is behaving differently: it’s all right with Dundag, it’s really the furthest from all other subdialects, but the Baļtinova-Vileks pair is not good – this distance turns out to be much bigger than distances from Vileks to Auleja and Rudzātys.

Mahalanobis distance is once again completely inane: the results are “hopping” in many ways and without any understandable explanation. Special Euclidean distance is equally seamless.

An amount of dialect spontaneous speech collected will never be exactly the same – it depends on circumstances, success, and other unpredictable factors during gathering. Of course, the amount of speech we collected also varies from subdialect to subdialect. It’s just normal, nothing wrong, but these differences in quantity can affect results of experiments: models trained on a significantly higher amount of speech may differ from others, just because the amount of speech is larger, and therefore differences based on linguistic characteristics, which are researchers interested in, can be lost and results distorted.

Therefore, we decided to carry out experiments in which the amount of speech is balanced, that is, for subdialects with a larger quantity of speech only in the initial part of each informant’s recording was left proportionally to the part we needed to minimize the amount, so the total amount of speech of each subdialect would be approximately the same as of the “smallest” one. Our hypothesis was that results should be better than with unbalanced speech data.

This – the **third** – experiment is the same as the first one, but “balanced” speech recordings are used instead of full ones.

As in the previous experiments, let’s start with **mean value vectors**. The results of Euclidean metrics in the balanced case do not differ significantly from the unbalanced one, obtained in the first experiment, that is, they are good: Dundag is the furthest for all other subdialects, Vileks and Baļtinova are the closest pair. As without balancing, Kullback-Leibler divergence contains negative values.

There are problems with city blocks metrics: Dundag appears to be closer than Rudzātys to Auleja, Baļtinova and Vileks. But without balancing, the L_1 metrics did not have such a problem. So it looks like the Euclidean metrics is more stable against balancing than the L_1 metrics. Jordan metrics of mean value vectors look just as meaningless as with unbalanced data: Dundag is closer to Vileks than all(!) other Latgalian subdialects, which also means that Baļtinova and Vileks are not the closest pair, and for Auleja Dundag appears to be closer than Vileks.

Euclidean metrics for **covariance matrices** behave similarly as without balancing, as well as similarly to mean vectors, i.e., no discrepancies are observed. The Kullback-Leibler divergence also behaves similarly to the unbalanced case; however, problems are in other areas: for Auleja and Baļtinova, Dundag turns out to be closer than Rudzātys. After symmetrysation, Auleja’s problem disappears, but Baļtinova’s remains.

City block metrics is more sensitive against balancing – it shows serious faults, though were not such problems before balancing: for Baļtinova Auleja is closer than Vileks and Dundag – closer that Rudzātys, but for Auleja Dundag is closer than both Rudzātys and Vileks. Also Jordan metrics, which before balancing did not had problems, is now showing them up, and many: for Auleja Dundag is closer than Rudzātys, for Baļtinova Vileks – further than Auleja, but for Vileks – Baļtinova further than Rudzātys.

In the case of the **weight vectors**, Euclidean metrics has the same faults as in the unbalanced case, however, they are less pronounced, that is, the difference has reduced: for Auleja Dundag is closer than Vileks, for Vileks Rudzātys is closer than Baļtinova.

The Kullback-Leibler divergence also has the same problems as in the unbalanced situation – some at the same place and some at other places: in one direction for Auleja Dundag is closer than Baļtinova and Vileks, for Baļtinova – Vileks further than Auleja and Dundag, for Vileks – Baļtinova further than Rudzātys; in other direction problems are similar, so, they are reciprocal. Therefore, it is clear that symmetrisation in this case cannot help and is not worth doing.

City blocks metrics in this case contains one problem less than it was without balancing. Only Auleja's mismatch remains, to which Dundaga appears closer than Baļtinova and Rudzātys. Unlike the previous one, Jordan metrics behave worse than without balancing: as if the problems are the same, but their number has doubled – not only Vileks appears to be very close to Dundag, but also Rudzātys.

Mahalanobis distance doesn't surprise us – it's still utterly pointless. And just like the special Euclidean distance.

On the other hand, the **fourth** experiment is the same as the second one, just "balanced" recordings are used instead of full-length ones.

Euclidean distance between **mean value vectors** behaves as without balancing, that is worse than without MAP adaptation: for Vileks and Baļtinova Dundag is closer than Auleja. Kullback-Leibler divergence is not understandable – it contains negative values.

City block metrics has the same fault as without balancing (and as Euclidean metrics), but to a lesser extent. Jordan metrics look more or less polite, but yet it has one other thing to blame: Dundag is much closer to Vileks than Rudzātys.

For **covariance matrices** Euclidean metrics behaves as in the case without balancing, i.e., it has the same problem as for mean vectors, but in a more pronounced form. Kullback-Leibler divergence is as much a fault as it was without balancing. And it's almost symmetrical, so obviously symmetrisation won't help.

City block metrics also behaves as without balancing, which means that it has the same problem as Euclidean distance, as well as an additional one: for Vileks Rudzātys is much closer than Baļtinova. Jordan metrics also show the same flaws as was for unbalanced data, but they are partly due to other subdialects and to a lesser extent.

Meanwhile, Euclidean metrics for **weight vectors** has been degraded as a result of the balancing: for Baļtinova now Dundag is closer than Rudzātys, but for Vileks – Rudzātys closer than Baļtinova. Kullback-Leibler divergence also seems to have only one grief, just as it does without balancing. The difference, however, lies in the fact that this time the affliction is symmetrical, so that it cannot be corrected by symmetrisation, and it must therefore be concluded that without balancing this divergence behaves better.

City blocks metrics behave as it does without balancing, that is, similar to Euclidean metrics, but worse. Jordan metrics behave worse than unbalanced: instead of one problem, it has a bunch of them – for Auleja and Baļtinova Dundag is closer than Rudzātys, also for Rudzātys Dundag is closer than Auleja, while for Baļtinova Auleja is closer than Vileks.

Mahalanobis distance also this time look completely pointless that becomes clear even seeing as how heavily all the distances differ from each other. Also, the so-called special Euclidean distance is not significantly better.

We can **conclude** that in general the method is applicable because it calculates data and uses metrics that give good results.

MAP adaptation, in general, makes results worse. This is what we could have imagined, because, in fact, adaptation is about facilitating data production process at the expense of differences between models, albeit to a small extent to make the recognition processes work. But what is good enough at speech recognition tasks may not be good at calculating distances.

The hypothesis of improving results as a result of data balancing has proved to be false – not only have the results not improved, but even slightly worsened. Perhaps because the fact that there were more data available for particular subdialects allowed these models to be developed in a better quality, thus also highlighting the objective differences between subdialects. It follows that we can expect better results in a case of increasing the amount of speech data.

If we look at the specific metrics and the data they are calculated for, then the Euclidean metrics for mean vectors and covariance matrices is good and stable against balancing. Both options are applicable, but it is possible that these two types of data can be combined to create some Euclidean metrics that take both of them into account.

Therefore, the main conclusion of this chapter is: Gaussian mixture models can be used to estimate a degree of proximity of languages, and it can be done with a simple and accessible Euclidean metrics.

Verification of the Results by Expert Evaluation Method

In all previous chapters, we used our knowledge from linguistically analytical sources and live communication with native speakers of subdialects to evaluate the newly developed methods. Such an assessment led us to believe that the methods are workable and usable, but formally, although based on objective sources, it may be considered subjective, that’s why it is also necessary to carry out a formal evaluation, preferably – numerical.

There are areas of computational linguistics where so-called “gold standards” are already developed. It is clear that there is no analysis of speech recordings of Latvian dialects among them. It would also be impossible to develop such a standard on their own, both in terms of labor and responsibility. Therefore, the examination of the results had to follow another path – a path of prospective comparative evaluation.

In this case, it seemed to us that the expert evaluation method is more accessible and feasible, both because other – recognised – numerical estimates have not been developed and because specialists needed to evaluate a proximity of Latvian dialects are available on the spot, namely in Latvia.

It is stated in the literature that the best specialists in the field should be selected as experts, and a teen number is reported as the optimum number of experts. As experiments were carried out on recordings of Latvian subdialects, a very small group of people who are or have been engaged in dialectology in Latvia were able to carry out this kind of manual evaluation. We identified all potentially competent people and asked almost everyone to evaluate our data. Therefore, we can conclude that we have made the optimal choice of experts – both in terms of qualitative and quantitative indicators.

All the methods we have developed are usable for calculation of distances between languages, and these numerical distances are also the output data of the methods. However, this type of results is too difficult to evaluate: in order to compare it with each other, we would have to standardize it in some (unknown to us) way, because the distances are relative, so each method is different in absolute terms, but experts would have undertake a numerical naming of such distances, which is unlikely to be possible, especially since the vast majority of professionals in the field are humanitarian-oriented.

In almost every chapter, we determined a success of a method by hierarchical clustering schemes, which were much more visually demonstrative than numerical distance tables. Apparently, through an expert survey, they should also be asked to draw trees of this kind of hierarchical clustering, by selecting the closest pair in each step, according to a professional knowledge and feelings.

First of all, we decided to look at certain known expert evaluation methods (EEM) and to assess their applicability to our data. They range from 4 to 7 in different sources, but they are all focused on arranging elements on a single axis, pairing, ranking or evaluating them in quantitative units, hence our results of a hierarchical clustering, which is expressed in binary trees, along with tags or labels on all leafs, but without tags on nodes, does not correspond to any of the methods given.

As our experiments were conducted on 3 different datasets – of 4 (non-parallel spontaneous speech transcriptions in phonetic transcription), 5 (spontaneous speech recordings) and 13 (parallel vocabularies) subdialects, we included 3 relevant tasks in the questionnaire and asked to hierarchically categorise each of the sets. We also added an additional page with an illustrative example for better understanding.

Since we have already selected the format of the comparable data – binary trees created by hierarchical clustering – we need to look for metrics defined in the space of such objects for further work. It should be noted that calculating distances between objects of this type is not a common task, but there have been some attempts to do so, so in the global context there are also offers of relevant metrics.

The most widespread is so-called rotation distance – defined as the minimum number of rotations needed to get another tree. There are also developed suggestions for computational algorithms. And even though it is basically working with trees without tags, it is argued that switching to tagged trees does not change the substance. However, how these rotations themselves are defined contradicts the nature of the objects we describe with binary trees: it turns out, for example, that a single rotation operation can produce (in our view) a fundamentally different tree that does not correspond to such a close (rotation) distance.

We found another, lesser known, AKM distance. In this, we were not satisfied with the very fact that in the lower branches differences are of lesser value, because in our case all choices are important, no matter when they are made.

As a result, our search concluded that none of the metrics found matched our task. To do this, on the other hand, we have to try to define ourselves a metric that quantifies the nature of the relationship between our objects.

What, then, is the difference that can be attributed to our evaluations, both automated and manual? In principle, these are choices that are made at each step by choosing the nearest pair. So, if a set of *n* languages is given, then the trees we create can also be described as *n*-elementary sets consisting of newly created (or selected) pairs, which at the next step become selection elements.

Let us call the **hierarchical choices distance** the division of volume of sets of all sets of non-trivial choices of hierarchical clustering by the number of non-trivial choices. The hierarchical choices distance is a metric.

In order for an expert evaluation to be usable, it must be ensured that they have a sufficiently high level of competence. There are several ways to do this, but for us, the only realistic option available was to measure competence on the basis of the survey data itself – also known as the degree of consensus, which by the words can be improved. Of course, there may be risks that most are incompetent, and they are the competents who are rejected, but we have reduced these risks by choosing the best professionals. Certainly, they are all described for standard estimation cases, but not for hierarchical clustering. Therefore, we had to try to come up with a similar experts' competence/consensus assessment that would fit our data.

Let us calculate distance of hierarchical choices between all the experts. For each expert, we will calculate the arithmetic mean distance to all other experts. Those whose average distance will be greater than one half (0.5), let's exclude and repeat all steps from the beginning with the remaining experts. When arithmetic mean distances between all remaining experts are less than 0.5, let's stop the process.

As the first set (4 subdialects) had the ratings coincided, we did not create a separate table for them because all the distances are equal to zero. When calculating mean values per set, this zero was taken into account as one of the three summands.

The results showed that the average distance of one expert is greater than 0.5, so we exclude it as not meeting the consensus criteria and recalculate the table without it in the next step.

In the next step, the consensus of experts is higher than in the first calculations, so the exclusion of the expert outside our limits has made a positive contribution. There are no more experts with an average distance greater than 0.5, so the expert competency assessment has been successfully completed and we can move on to the main task of evaluating the methods.

Next, we calculated the distances between all the methods and all the experts, with and without exclusion of the ineligible expert (to see how it affects results). The limits of the usability of the methods must be intuitively determined by us ourselves.

We did this in two ways – relative (the limit is the average distance of experts for a given set, i.e. method's error should not exceed the error of experts) and absolute (the limit is an absolute number set by us, which is based on intuition and experience). The results show that there are enough methods that are considered to be good after using the expert evaluation method.

Also, there is no ready recipe for error estimation when measurements are binary trees. In essence, the average value of the expert evaluation can be considered as the error of the method – the higher it is, the worse the method is, i.e., the more erroneous. Since such an approach is empirically consistent and does not require additional work investment, we have also followed it.

Firstly, with the expert method, we have formally demonstrated the viability of our automated methods and evaluated the error of each method. One did not have it at all, some – had a small (these methods are certainly usable), few – a medium (possibly usable), others – too large (unusable).

Secondly, it should be concluded that the expert method in hierarchical clustering is applicable to relatively small sets (up to 10?) of elements (languages) – with 13 elements it is already difficult for a person to orientate and the spontaneity of decisions begins to prevail over reasonableness.

Thirdly, if an automated evaluation is done on basis of certain characteristics (e.g. phonetic, morphological, lexical or syntactic), then an expert evaluation must be done separately by the same characteristics. Otherwise, making one single assessment can increase the error significantly (and even critically).

Afterwords

As we have verified during the dissertation experiments, our main hypothesis is confirmed: both phonetic transcription of speech and recordings of speech as input data are sufficient to derive numerical estimation of a degree of proximity of languages by statistical methods.

The goal of the work has been achieved: we have developed six new methods that meet the requirements we have defined (see preface). All four tasks have also been completed.

Because of the limited volume of phonetically transcribed speech and varying traditions of transcription, which require quite a lot of work to unify, it is obviously more promising to work with speech recordings. At the same time, the techniques for phonetic transcription are more visual and elegant and should not be neglected. In addition, as we have shown in the chapter of the method of phoneme recognisers, there is also a potential to merge these seemingly completely different approaches.

The most recent and easy-to-use method for recording speech recordings is the i-vector method. While building a complex system for determining a degree of language proximity, this method, probably, should be taken as a basis, but it should be complemented by other methods too. Such a system could be used to achieve the goal that induced us to start working on this topic: to restrict public and political speculations on the issue where a dialect ends and a language begins.

We are currently working on the extraction and preparation of larger volumes of material (which is a large and voluminous job). We want to apply our methods to a larger number of languages so that the categorisation results reflect the closeness of all languages within a whole language family (or multiple families), so comparing the results to the analytical language breakdown will be more interesting and more likely to evaluate.

As part of our internship at the Lithuanian Language Institute (Vilnius), we have prepared a corpus of all Lithuanian subdialects, for which we will try to apply our methods. Similarly, during the internship at the National Autonomous University of Mexico, we began work on a speech corpus of Spanish dialects and a speech corpus of Indian languages, for which it will also be possible to apply the methods described in this dissertation.

In the long run, we would gladly see a super-method based on our certain methods, which would combine all available particular methods and calculate a distance between any language pair. The software should have a convenient interface, polished, widely used. The calculated distances would be internationally recognised and taken into account in both scientific and public opinion. All the languages of the world would have enough input to use this super-method. It certainly isn’t a work of one person, but of several institutes. But we have made at least the core – by opening this door and proving that technology can be created and applied.

References

We created the references in chronological order – just as the sources "came" to us during writing the dissertation. This sequence is different from the dissertation sequence because its chapters were written in mixed order.

We tried to be as honest as possible in describing of literature were used – we identified all the sources we used in our cognitive process, including websites and web encyclopedias. We believe that honesty and openness must be established in the first place in science, and if we use a source, it should be stated. On potentially disdainful comments, including about the use of Wikipedia, we explain that Wikipedia is a usable and even a useful source, provided of course that it is handled properly.

1. *Никольский В.К., Яковлев Н.Ф.* Как возникла человеческая речь. Москва: Государственное издательство культурно-просветительной литературы, 1949. 64 стр. с ил.
http://genling.ru/books/item/f00/s00/z0000029/index.shtml
Piekļūts 2015. gada 1. aprīlī.

2. *Ferdinand de Saussure.* Cours de linguistique générale. Paris: Payot, coll. «Grande bibliothèque Payot », 1995 (1re éd. 1916).
http://fr.wikisource.org/wiki/Page:Saussure_-_Cours_de_linguistique_g%C3%A9n%C3%A9rale,_%C3%A9d._Bally_et_Sechehaye,_1971.djvu/32
Piekļūts 2015. gada 1. aprīlī.

3. *Иванов В.В.* Моногенеза теория // Лингвистический энциклопедический словарь. - М., 1990. - С. 308-309.
http://www.philology.ru/linguistics1/ivanov-90c.htm
Piekļūts 2015. gada 1. aprīlī.

4. *Кочеткова В.И.* Палеоневрология. М.: Изд-во Моск. ун-та, 1973. С. 188-215.
http://www.ido.rudn.ru/psychology/anthropology/ch4_2.html
Piekļūts 2015. gada 1. aprīlī.

5. *Кареев Н.И.* О «новом взгляде» г. Шапиро на современную систему сравнительного языкознания (Возражение) // Филологические записки. Воронеж, 1874.
http://vrn-id.ru/filzaps741.htm
Piekļūts 2015. gada 1. aprīlī.

6. *Топоров В.Н.* Сравнительно-историческое языкознание // Лингвистический энциклопедический словарь. - М., 1990. - С. 486-490.
http://www.philology.ru/linguistics1/toporov-90.htm
Piekļūts 2015. gada 1. aprīlī.

7. *Иванов В.В.* Генеалогическая классификация языков // Лингвистический энциклопедический словарь. - М., 1990. - С. 93-98.
http://tapemark.narod.ru/les/093d.html
Piekļūts 2015. gada 1. aprīlī.

8. *Яхонтов С.Е.* Оценка степени близости родственных языков // Теоретические основы классификации языков мира. - М., 1980. - С. 148-157

9. *Swadesh M.* Perspectives and problems of Amerindian comparative linguistics // Word, 1954, № 10, pp. 306-332.

10. *Дьячок М.Т., Шаповал В.В.* Генеалогическая классификация языков. - Новосибирск, 2002. – 32 с.
http://www.philology.ru/linguistics1/dyachok-shapoval-02.htm
Piekļūts 2015. gada 1. aprīlī.

11. *Реформатский А.А.* Введение в языкознание. М.: ГУПИ МП РСФСР, 1960.
431 стр.

12. דער ייִוואָ און די פּראָבלעמען פֿון אונדזער צײַט. *מ. ווײַנרײַך* // Yivo Bleter. – New York, 1945. Vol. XXV, No.3, p. 3-18.

13. *Зиндер Л.Р.* Общая фонетика. М.: Высшая школа, 1979. – 312 с.

14. Latviešu valodas dialektoloģijas atlanta materialu vākšanas programa. Rīga: Latvijas PSR ZA izdevniecība, 1954.

15. *Rudzīte M.* Latviešu dialektoloģija. Rīga: Latvijas valsts izdevniecība, 1964.

16. *Tambovtsev Y.* Phonological Similarity Between Basque and Other World Languages Based on the Frequency of Occurrence of Certain Typological Consonantal Features // The Prague Bulletin of Mathematical Linguistics 79-80, pp. 121-126, 2003.

17. *Canvar W.B., Trenkle J.M.* N-Gram-Based Text Categorization // In Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, NV, UNLV Publications/Reprographics, pp. 161-175, 1994.

18. Latviešu izlokšņu teksti. Sast. Marta Rudzīte. Rīga: P. Stučkas Latvijas Valsts universitāte, 1963.

19. Augšzemnieku dialekta teksti. Latgaliskās izloksnes. Sast. N. Jokubauska. Rīga: Zinātne, 1983.

20. Latviešu izlokšņu teksti. Sast. Benita Laumane. Liepāja: Liepājas Pedagoģiskā akadēmija, 2000.

21. *Ball M.J.* Teaching Vowels in Practical Phonetics: The Auditory or Articulatory Route?
http://www.phon.ucl.ac.uk/home/johnm/ball.htm
Piekļūts 2007. gada 26. oktobrī.

22. Corporate Author International Phonetic Association. Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet. Cambridge: Cambridge University Press, 1999.

23. *Kessler B.* Computational dialectology in Irish Gaelic // Proceedings of the European ACL. Dublin, 1995. Pp. 60-66.

24. *Markus D., Grigorjevs J.* Fonētikas pētīšanas un vizualizēšanas metodes, II grām. sēr.: Fonētikas pētīšanas un vizualizēšanas metodes. Rīga: Rasa ABC, 2004.

25. *University College London, Dept. of Phonetics and Linguistics.* Cardinal Vowels by Daniel Jones. London: UCL Press, 1996.

26. *Пиотровский Р.Г.* Еще раз о дифференциальных признаках фонемы // Вопросы языкознания, № 6, М.: РАН, 1960, стр. 24-38.

27. *Nerbonne J., Heeringa W., van den Hout E., van der Kooi P., Otten S., van de Vis S.W.* Phonetic Distance between Dutch Dialects // CLIN VI, Papers from the sixth CLIN meeting. Antwerp: University of Antwerp, Center for Dutch Language and Speech. Pp. 185-202.

28. От аналоговой записи – к цифре.
http://its-journalist.ru/Articles/ ot_analogovoj_zapisi_k_cifre.html
Piekļūts 2012. gada 5. martā.

29. Звукозапись, цифровая или аналоговая?
http://www.midi.ru/forumd.php?id=181648
Piekļūts 2012. gada 5. martā.

30. *Дубровский Д.Ю.* Чем цифровая запись лучше аналоговой?
http://demorecord.ru/analogsound.html
Piekļūts 2012. gada 5. martā.

31. *Музыченко Е.В.* Принципы цифрового звука. 1998-1999.
http://www.websound.ru/articles/theory/digsnd.htm
Piekļūts 2012. gada 5. martā.

32. Audio file format.
http://en.wikipedia.org/wiki/Audio_file_format
Piekļūts 2012. gada 5. martā.

33. Сжатие без потерь.
http://ru.wikipedia.org/wiki/Сжатие_без_потерь
Piekļūts 2012. gada 5. martā.

34. Сжатие данных с потерями.
http://ru.wikipedia.org/wiki/Сжатие_данных_с_потерями
Piekļūts 2012. gada 5. martā.

35. A-Law Compressed Sound Format.
http://www.digitalpreservation.gov/formats/fdd/fdd000038.shtml
Piekļūts 2012. gada 5. martā.

36. *Salvi G.* Mining Speech Sounds. Stockholm: KTH, 2006. Pp. 18-19.

37. *Melin H.* Automatic speaker verification on site and by telephone: methods, applications and assessment. Stockholm: KTH, 2006. Pp. 103-104.

38. Микрофон.
http://ru.wikipedia.org/wiki/Микрофон
Piekļūts 2012. gada 5. martā.

39. Электретный микрофон.
http://ru.wikipedia.org/wiki/Электретный_микрофон
Piekļūts 2012. gada 5. martā.

40. Характеристики микрофонов.
http://ingibit.rigalink.lv/info/c2/mikro01.html
Piekļūts 2012. gada 5. martā.

41. Сравнение конденсаторных и динамических микрофонов.
http://www.microphone.ru/articles/paragraph_1.html
Piekļūts 2012. gada 5. martā.

42. Динамические, конденсаторные микрофоны и фантомное питание.
http://midi.ucoz.ru/publ/1-1-0-16
Piekļūts 2012. gada 5. martā.

43. Конденсаторный микрофон.
http://ru.wikipedia.org/wiki/Конденсаторный_микрофон
Piekļūts 2012. gada 5. martā.

44. *Костоломов В.* Ленточные микрофоны. 2000.
http://www.oktava-mics.net/shop/a-2/lentochnye_mikrophony.html
Piekļūts 2012. gada 5. martā.

45. Угольный микрофон.
http://ru.wikipedia.org/wiki/Угольный_микрофон
Piekļūts 2012. gada 5. martā.

46. *Rabiner L.* A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition // Proceedings of the IEEE, vol. 77, No 2, February 1989, p. 262-286.

47. *Кушнир Д.А.* Алгоритм формирования структуры эталона для пословного дикторонезависимого распознавания команд ограниченного словаря // Штучный інтелект, № 3'2006. Київ, 2006.

48. Dainuskapis. Sast. Kr. Barons.

49. *Haugland Tokheim Å.E.* iVector Based Language Recognition. Trondheim: NTNU, 2012.

50. *Li H., Ma B., Lee K.A.* Spoken Language Recognition: From Fundamentals to Practice // Proceedings of the IEEE, Vol. 101, No. 5, May 2013.

51. *Plchot O., Diez M., Soufifar M., Burget L.* PLLR Features in Language Recognition System for RATS // Interspeech. Singapore, 2014. Pp. 3047-3051.

52. *Tebelskis J.* Speech Recognition using Neural Networks. Pittsburgh: Carnegie Mellon University, 1995.

53. *Liu Z., Huang Q.* A new distance measure for probability distribution function of mixture type // IEEE International Conference on Acoustics, Speech, and Signal Processing, 2000. Vol. 1, pp. 616-619.

54. *Young S., Evermann G., Gales M., Hain Th., Liu X., Moore G., Odell J., Ollason D., Povey D., Valtchev V., Woodland Ph.* The HTK Book (for HTK Version 3.4). Cambridge: Cambridge University Engineering Department, 2009.

55. *Johnson D., Sinanovic S.* Symmetrizing the Kullback-Leibler Distance. Computer and Information Technology Institute, Department of Electrical and Computer Engineering, Rice University, Houston, 2001.

56. Метрика // Математическая энциклопедия. – М.: Советская энциклопедия, 1982. – Т. 3.

57. *Howard D., Angus J.* Acoustics and psychoacoustics. Oxford: Focal Press, 2009.

58. *Kullback S., Leibler R.* On information and sufficiency // Annals of Mathematical Statistics. Vol. 22, No. 1, Mar 1951, pp. 79-86.

59. *Dehak N., Dehak R., Kenny P., Brummer N., Ouellet P., Dumouchel P.*. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification // Proceedings Interspeech. Brigthon, UK, 2009.

60. *Dehak N., Kenny P.J., Dehak R., Dumouchel P., Ouellet P.* Front-End Factor Analysis for Speaker Verification // IEEEE Transactions On Audio, Speech, And Language Processing. Piscataway: IEEE Press, 2011. Vol. 19, no. 4, pp. 788-798.

61. *Schwarz P.* Phoneme Recognition based on Long Temporal Context, PhD Thesis. Brno: Vysoké učení technické v Brně, 2009.

62. *Schwarz P., Matejka P., Cernocky J., Chytil P.* Phonotactic Language Identification using High Quality Phoneme Recognition // Proceedings Eurospeech, 2005.

63. Phoneme Recognition (caveat emptor) // CMU Sphinx.
http://cmusphinx.sourceforge.net/wiki/phonemerecognition
Piekļūts 2016. gada 16. februārī.

64. Phoneme recognizer based on long temporal context // BUT Speech@FIT.
http://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context
Piekļūts 2016. gada 16. februārī.

65. *Lamere P., Kwok P., Walker W., Gouvea E., Singh R., Raj B., Wolf P.* Design of the CMU Sphinx-4 decoder // Proceedings of the 8th European Conference on Speech Communication and Technology, Geneve, Switzerland, 2003, pp. 1181–1184.

66. *Soufifar M.* Subspace Modeling of Discrete Features for Language Recognition. Trondheim: NTNU, 2014.

67. *Ghosh S., Vijay Girish K.V., Sreenivas T.V.* Relationship between Indian Languages Using Long Distance Bigram Language Models // Proceedings of ICON-2011: 9'th International Conference on Natural Language Processing. Chennai: Macmillan Publishers, 2011.

68. *Zha Sh., Peng X., Cao H., Zhuang X., Natarajan P., Natarajan P.* Text Classification via iVector Based Feature Representation // 2014 11th IAPR International Workshop on Document Analysis Systems. IEEE, 2014.

69. *Dehak N., Torres-Carrasquillo P.A., Reynolds D., Dehak R.* Language Recognition via Ivectors and Dimensionality Reduction // Proceedings of Interspeech 2011. Florence: International Speech Communication Association, 2011.

70. *Glembek O., Burget L., Matejka P.* Voice Biometry Standard – Draft. Brno: Speech@FIT, 2015.

71. *Nouri J., Yangarber R.* Measuring Language Closeness by Modeling Regularity // Proceedings of the EMNLP’2014 Workshop: Language Technology for Closely Related Languages and Language Variants. Doha: 2014.

72. Вавилонская Башня: Проект этимологической базы данных.
http://starling.rinet.ru/
Piekļūts 2016. gada 7. martā.

73. *Cilibrasi R., Vitányi P.M.B.* Clustering by compression // IEEE Transactions on Information Theory. Toronto: IEEE, 2005.

74. *Miao Y., Zhang H., Metze F.* Towards speaker adaptive training of deep neural network accoustic models // Proceedings of 15th Annual Conference International Speech Community Associtation. Singapore: Interspeech, 2014.

75. *Yao K., Yu D., Seide F., Su H., Deng L., Gong Y.* Adaptation of context-dependent deep neural networks for automatic speech recognition // Proceedings of the Spoken Language Technology Workshop. Miami: SLT, 2012.

76. *Saon G., Soltau H., Nahamoo D., Picheny M.* Speaker Adaptation of Neural Network Acoustic Models Using I-Vectors // Proceedings of Automatic Speech Recognition and Understanding (ASRU) Workshop. Olomouc: IEEE, 2013.

77. *Trumpa E.* Latviešu ģeolingvistikas etīdes. R.: Zinātne, 2012.

78. *Садыхов Р.Х., Ракуш В.В.* Модели гауссовых смесей для верификации диктора по произвольной речи // Доклады БГУИР, № 4/2003. Минск: Белорусский государственный университет информатики и радиоэлектроники, 2003.

79. *Dobkeviča M.* Varbūtību teorijas un matemātiskās statistikas elementi. Daugavpils: RTU DF, 2004.

80. *Davis S., Mermelstein P.* Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences // IEEE Transactions on Acoustics, Speech, and Signal Processing, 1980, Vol. 28, No. 4, pp. 357-366.

81. *Reynolds D.* Universal Background Models. MIT Lincoln Laboratory.
https://www.ll.mit.edu/mission/cybersec/publications/publication-files/full_papers/0802_Reynolds_Biometrics_UBM.pdf
Piekļūts 2016. gada 21. septembrī.

82. *Reynolds D., Quatieri T., Dunn R.* Speaker Verification Using Adapted Gaussian Mixture Models // Digital Signal Processing 10, pp. 19-41, 2000.

83. *Smyth P.* The EM Algorithm for Gaussian Mixtures. // Probabilistic Learning: Theory and Algorithms. Irvine: University of California.
http://www.ics.uci.edu/~smyth/courses/cs274/notes/EMnotes.pdf
Piekļūts 2016. gada 21. septembrī.

84. *Reynolds D.* Gaussian Mixture Models. MIT Lincoln Laboratory.
https://www.ll.mit.edu/mission/cybersec/publications/publication-files/full_papers/0802_Reynolds_Biometrics-GMM.pdf
Piekļūts 2016. gada 21. septembrī.

85. *Chang W., Cathcart Ch., Hall D., Garrett A.* Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis // Language, 2015, Vol. 91, No. 1, pp. 194-244.

86. *Hennig W.* Grundzüge einer Theorie der phylogenetischen. Berlin: Deutscher Zentralverlag, 1950.

87. *Hennig W.* Phylogenetic Systematics // Annual Review of Entomology, 1965, Vol. 10, pp. 97-116.

88. *Булатова Л.Н., Касаткин Л.Л., Строганова Т.Ю.* О русских народных говорах. М.: Просвещение, 1975.

89. *Zinkevičius Z.* Lietuvių kalbos tarmės. Kaunas: Šviesa, 1968.

90. *Chapman W.H., Olsen E., Lowe I., Andersson G.* Introduction to Practical Phonetics. Horsleys Green: Summer Institute of Linguistics, 1989.

91. *Кочерган М.П.* Вступ до мовознавства: Підручник для студентів філологічних спеціальностей вищих закладів освіти. - К.: Видавничий центр Академія, 2001.

92. *Scupin R.* Cultural Anthropology: A Global Perspective. Boston: Pearson, 2012.

93. *Левенштейн В.И.* Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады Академии наук СССР. Том 163, вып. 4, стр. 845-848. М.: Наука, 1965.

94. *Винцюк Т.К.* Распознавание слов устной речи методами динамического программирования // Кибернетика. Вып. 1, стр. 81-88. Киев: Наукова думка, 1968.

95. *Wagner R.A., Fischer M.J.* The string to string correction problem // Journal of the ACM. New York: ACM, 1974. Vol. 21, no. 1, pp. 168–173.

96. *Kaufman L., Rousseeuw P. J.* Finding Groups in Data – An Introduction to Cluster Analysis. New Jersey: Wiley, 1990.

97. *Van Noord G.* TextCat.
http://www.let.rug.nl/~vannoord/TextCat/
Piekļūts 2004. gada 10. oktobrī.

98. *Han J., Kamber M., Pei J.* Data Mining: Concepts and Techniques. 3rd Edition. Waltham: Morgan Kaufmann / Elsevier, 2012.

99. *Drgas Sz., Dąbrowski A.* Generalized cosine similarity in I-vector based automatic speaker recognition systems // Signal Processing: Algorithms, Architectures, Arrangements, and Applications. Poznan: IEEE, 2013. Pp. 73-77.

100. *Bai Zh., Zhang X.-L., Chen J.* Cosine Metric Learning for Speaker Verification in the i-Vector Space // Interspeech 2018. Hyderabad: 2018. Pp. 1126-1130.

101. *Берзиньш А.У.* Сравнение балтийских языков методом n-грамм // Труды международной коференции «Корпусная лингвистика - 2004». СПб.: Издательство С.-Петербургского университета, 2004.

102. *Berzinch A.A.* La comparaison de typologie traditionnelle et de typologie phonolexique, basée sur la méthode des n-grammes, dans les dialectes baltes // Identification des langues et des

variétés dialectales par les humains et par les machines. Paris: École National Supérieure des Télécommunications, 2004.

103. *Берзинь А.У.* Измерение фономорфолексического расстояния между латышскими наречиями путём применения расстояния Вагнера-Фишера // Труды международной конференции «Диалог 2006». М.: Издательство РГГУ, 2006.

104. *Bērziņš A.A., Grigorjevs J.* Latviešu izloksnēs sastopamo fonēmu telpa // Linguistica Lettica XVIII, R.: Latviešu valodas institūts, 2008.

105. *Берзинь А.* Возможности применения статистических методов распознавания речи для определения близости языков // Прикладна лінгвістика та лінгвістичні технології, Megaling-2009. Київ: «Довіра», 2009.

106. *ბერზინი ა.* ინფორმაციის მოპოვების პრინციპები ფონოგრამების ავტომატური ანალიზისთვის / Принципы сбора информации для автоматизированного анализа фонограмм // ქართული ენა და თანამედროვე ტექნოლოგიები - 2011. თბილისი: „მერიდიანი“, 2011.

107. *Берзинь А.У.* Применение распознавателей фонем для автоматического определения уровня близости языков // Труды международной конференции «Диалог 2016». М., 2016.

108. *Bērziņš A.A.* Usage of HMM-based Speech Recognition Methods for Automated Determination of a Similarity Level between Languages // AINL Proceedings. «Springer», 2019.

109. *Berziñch A.A., Chavarría Amezcua M.A.* El espacio de los alófonos del español. Iesniegts publicēšanai «Lengua y Habla», 2020.

110. *Павлинов И.Я.* Введение в современную филогенетику (кладогенетический аспект). М.: изд-во КМК, 2005.

111. *Šimko J., Suni A., Hiovain K., Vainio M.* Comparing Languages Using Hierarchical Prosodic Analysis // Proceedings Interspeech 2017. Stockholm: 2017. Pp. 1213-1217.

112. Ekonomikas skaidrojošā vārdnīca. *Sast. aut. kol. R. Grēviņas vadībā.* R.: Zinātne, 2000.

113. Valodniecības pamatterminu skaidrojošā vārdnīca. *Atb. red. V. Skujiņa.* R.: LU Latviešu valodas institūts, 2007.

114. LVS ISO 5127:2005. Informācija un dokumentācija. Vārdnīca. Informācijas zinātnes termini (bibliotēkas, arhīvi un muzeji). R.: 2005.

115. Естественный язык.
http://ru.wikipedia.org/wiki/Естественный_язык
Piekļūts 2019. gada 14. septembrī.

116. *Gay K.M.* Recent Advances and Issues in Computers. Phoenix, Arizona: Oryx Press, 2000.

117. Demogrāfija 2018, statistisko datu krājums. R.: Centrālā statistikas pārvalde, 2018.

118. *Mehl M.R., Vazire S., Ramírez-Esparza N., Slatcher R.B., Pennebaker J.W.* Are Women Really More Talkative Than Men? // Science, No. 317 (5832). Washington: American Association for the Advancement of Science, 2007.

119. *Liberman M.* Sex-Linked Lexical Budgets // Language Log. 2006/2007.
http://itre.cis.upenn.edu/~myl/languagelog/archives/003420.html
Piekļūts 2019. gada 15. septembrī.

120. *Čmejrková S.* The (Re)Presentation Of The Author In Czech And Slovak Scientific Texts // Jezik in slovstvo, Vol. 52 , Issue 3–4. Ljubljana: Zveza društev Slavistično društvo Slovenije, 2007.

121. *Жеребило Т.В.* Словарь лингвистических терминов: Изд. 5-е, испр. и дополн. — Назрань: Изд-во «Пилигрим», 2010.

122. *Sarkar A., Matrouf D., Bousquet P.-M., Bonastre J.-F.* Study of the Effect of I-vector Modeling on Short and Mismatch Utterance Duration for Speaker Verification // Interspeech Proceedings. Portland, 2010.

123. *Громова Н.М., Громова Н.И.* Основы экономического прогнозирования. Учебное пособие. Старая Русса: Старорусский политехнический колледж, 2007.

124. *Markovičs Z.* Ekspertu novērtējuma metodes. R.: RTU izdevniecība, 2009.

125. Экспертное оценивание.
http://ru.wikipedia.org/wiki/Экспертное_оценивание
Piekļūts 2019. gada 31. oktobrī.

126. *Бешелев С.Д., Гурвич Ф.Г.* Экспертные оценки. М.: «Наука», 1973.

127. *Sleator D.D., Tarjan R.E., Thurston W.P.* Rotation Distance, Triangulations, and Hyperbolic Geometry // Journal Of The American Mathematical Society. Volume 1. Number 3. July 1988.

128. *Duda J.* Practical estimation of rotation distance and induced partial order for binary trees. Cornell University, 2016.

129. *Chen Y.J., Chang J.M., Wang Y.L.* An efficient algorithm for estimating rotation distance between two binary trees // International Journal of Computer Mathematics.
Vol. 82, No. 9, September 2005, Taylor & Francis.

130. *Dehornoy P.* On the rotation distance between binary trees // Advances in Mathematics. No. 223. Elsewier: 2010.

131. *Caspersen K.M., Madsen M.B., Eriksen A.B., Thiesson B.* A Hierarchical Tree Distance Measure for Classification // Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods. Porto: “SCITEPRESS”, 2017.

132. *Полегенько А.Ф., Князский О.В.* Оценка относительной компетентности экспертов в экспертной группе с использованием матриц парных сравнений // Озброєння та військова техніка. № 3. Київ: Центр. НДІ озброєння та військ. техніки ЗС України, 2014.

133. *Бурков Е.А.* Определение компетентности экспертов на основе поставленных ими оценок // Известия СПбГЭТУ «ЛЭТИ». № 4. Санкт-Петербург, 2009.

134. *Берзинь А.У.* Применение i-векторов для автоматизированного определения уровня близости языков // Труды Института системного программирования РАН. Том 31, № 5. М.: ИСП РАН, 2019.

135. *Bērziņš A.A.* Automated Comparison of Natural Languages – Software and Datasets of the Dissertation (Thesis). "Zenodo", 2019.
http://doi.org/10.5281/zenodo.3527981
Piekļūts 2019. gada 4. novembrī.

136. Preliminary recommendations on Corpus Typology. EAGLES – Expert Advisory Group on Language Engineering Standards Guidelines, 1996.
http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html
Piekļūts 2019. gada 5. novembrī.

137. *Maia B.* What are comparable corpora? // Proceedings of the workshop on Multilingual Corpora: Linguistic Requirements and Technical Perspectives. Lancaster: Saarland University, 2003.

138. Comparable Corpora // MT Research Survey Wiki. University of Edinburgh.
http://www.statmt.org/survey/Topic/ComparableCorpora
Piekļūts 2019. gada 5. novembrī.

139. Similarity (State of the art) // ACL Wiki for Computational Linguistics. The Association for Computational Linguistics.
https://aclweb.org/aclwiki/Similarity_(State_of_the_art)
Piekļūts 2019. gada 6. novembrī.

140. *Nissani M.* Fruits, Salads, and Smoothies: A Working Definition of Interdisciplinarity // The Journal of Educational Thought (JET) / Revue de la Pensée Éducative. Vol. 29, No. 2. Calgary: Werklund School of Education, University of Calgary, 1995.

141. EURAB report makes recommendations to promote interdisciplinary research // CORDIS, EU research results. European Commission, 2004.
https://cordis.europa.eu/article/rcn/21983/en
Piekļūts 2019. gada 6. novembrī.

142. Par prioritārajiem virzieniem zinātnē 2018.–2021. gadā. Kopsavilkums. Izglītības un zinātnes ministrijas Augstākās izglītības, zinātnes un inovāciju departaments, 2017.

143. *Тимофеев К.А.* Религиозная лексика русского языка как выражение христианского мировоззрения: учебное пособие. Новосибирск, 2001.

144. Muzeoloģijas terminu vārdnīca. R.: Latvijas Muzeju asociācija, 1997.

145. *Balodis M.* Kurzemes cietoksnis // Austrālijas latvietis. Nr. 2855. Haknija, 25.VII.2007.

146. *Gārša A.* Minoritātes Latvijā vēsturiskā skatījumā // Brīvā Latvija. Nr. 22. 11.VI.2011.

147. Solomon Kullback.
https://en.wikipedia.org/wiki/Solomon_Kullback
Piekļūts 2020. gada 8. februārī.

148. *Даль В.И.* Толковый словарь живаго великорускаго языка. Томъ второй. И – О. С.-Петербургъ/Москва: Изданіе книгопродавца-типографа М. О. Вольфа, 1881.

149. Moisejs Kuļbaks.
https://timenote.info/lv/Moisejs-Kulbaks
Piekļūts 2020. gada 8. februārī.

150. *Passricha V., Aggarwal R.K.* Convolutional Neural Networks for Raw Speech Recognition // From Natural to Artificial Intelligence: Algorithms and Applications. IntechOpen, 2018.

151. *Dauphin Y.N., Fan A., Auli M., Grangier D.* Language Modeling with Gated Convolutional Networks // Proceedings of the 34th International Conference on Machine Learning. Volume 70, August 2017.