Using Machine Translation to Enable Highly Multilingual Scholarship

Daniel Ross, Ph.D. Candidate, University of Illinois

First Machine Translation Marathon in the Americas

May 13th, 2015

Early dreams for MT realized

- Due to funding sources and the interests of developers, an early goal of MT since the 1950s and earlier was to enable access to academic material written in other languages.
- Early research became a spectacle in the media
- They said MT would be ready in 5 years.
- …not quite. But 60 years later, MT, with domain-expert post-editing is capable of fulfilling this dream.

[From *Globe and Mail* (Toronto, 12 February 1954]

Electronic Translations Possible in Five Years

New York, Feb. 11 (UP).—Aan electronic brain smart enough to translate a whole Russian book at a glance may be developed within the next five years, a mathematican said today.

Such a robot scholar might even be educated well enough, he said, to pick out only the interesting parts of any foreign-language publications for conversion to English.

Peter Sheridan, an International Business Machines Corporation mathematician, said if scientists started working today they could perfect such a linguistic machine in three to five years.

Whether IBM is working on such a gadget is classified information, Sheridan said. But he indicated that Georgetown University experts may have started, work on such a project.

Under a Georgetown assignment, Sheridan has spent the last six months feeding Russian into IBM's newest and most-vocabulary of 250 words and eventually it translated 200 sentences. It would be possible for 701 to "learn" about 500 words, but another machine will have to be developed for further linguistic work.

"With a 'dictionary' of 60,000 entries and 60 to 100 rules of syntax, a computer could take a whole area of technical literature and translate freely," he said.

"It would be most useful in libraries where research could be carried out by scholars without their having to take time to do the translating."

IBM has rented several of its 701 models to the Government and industry for important mathematical computations. It is manufacturing about 18 of the brains each year.

http://www.mt-archive.info/GlobeMail-1954.pdf

Post-editing by domain experts

- Schwartz (2014) and Schwartz et al. (2014) showed that post-editing by domain experts can rescue imperfect output of current MT technologies.
- Domain experts: for example, academics who have a strong background in their fields
- Post-editing: the inspection and correction of MT output, with or without reference to the original text
- In this way, MT can provide effective access to material.

MT still needs improvement

- Languages supported
- Amount of training data
- More accessible and customizable interfaces

Possible applications

- Verifying relevance of a source
- Finding the right section to translate
- Help in translating a language of limited familiarity
- Full translation of unfamiliar language

Topics covered today

- Linguistic diversity and expectations
- Scanning
- Optical Character Recognition (OCR)
  - Specifically, Adobe Acrobat and Abbyy Finereader
- Machine Translation (MT)
  - Specifically, Google Translate for demo
- Post-editing

My background

- Ph.D. candidate at UIUC
- Expertise/experience in theoretical syntax, historical linguistics, typology/comparative linguistics.
- Studied 20 languages: Spanish, Italian, German, Latin, Arabic, Japanese, Portuguese, Hindi, French, Swahili, Catalan, Swedish, Basque, Modern Greek, Turkish, Mandarin Chinese, American Sign Language, Faroese, Quechua and Russian. (I don't *speak* all of these!)
- Citing sources in 30+ languages for my dissertation, using the methods described today.

Linguistic diversity

- Languages vary in possibly unexpected ways
  - Can we effectively expected the unexpected?
- Basic typological properties can help
- Many languages used in research are somewhat familiar and relatively similar to English
- Examples to consider:
  - Word order variation
  - Morphological type

Possible Word Orders

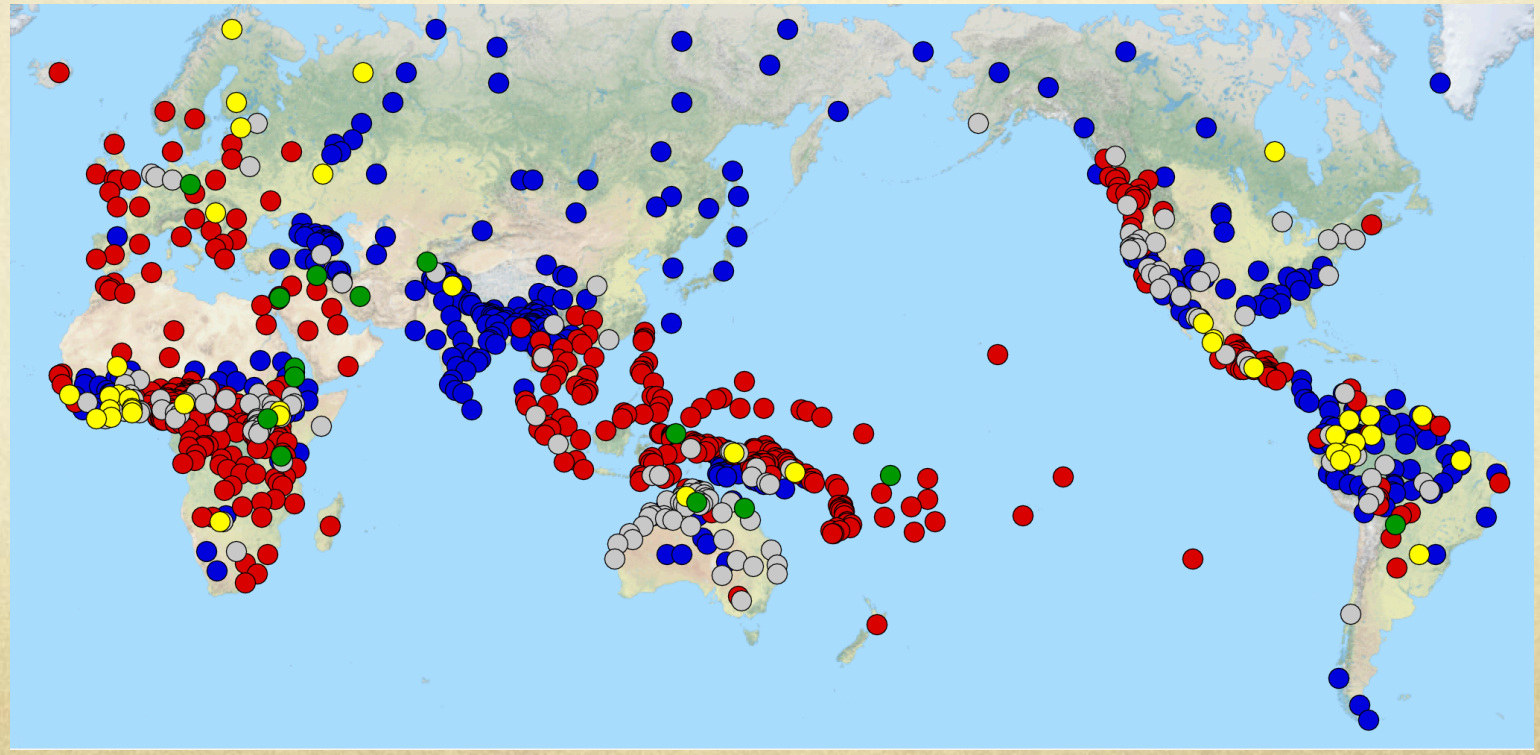
- ‘The boy loves the girl’ (examples in Latin)

Puer amat puelam	SVO
Puer puelam amat	**SOV**
Amat puer puelam	VSO
Amat puelam puer	VOS
Puelam puer amat	OSV
Puelam amat puer	OVS

- ‘The girl loves the boy’ Puela amat puerum

World Atlas of Language Structures

- Chapter 95a: V&O and Adpositions (wals.info)
  - Red: VO+PrePs; Blue: OV+PostPs (expected)
  - Yellow: VO+PostPs; Green: OV+PrePs; Gray: other



Morphological type

- Fusional/inflectional: Latin, *-us* may encode nominative, singular, masculine, 2nd declension class…
- Isolating/analytic: Chinese, no/limited morphology
- Agglutinative: Turkish & Swahili, many morphemes "glued" together, little overlap in function
- What about polysynthetic languages?
  - Untussuqatarniksaitengqiggtuq (Yupik; Payne 1997: 28)

"He had not yet said again that he was going to hunt reindeer."

Expected errors

- Knowing a little bit about linguistic typology and properties of the source language can go a long way.
- Does the language have cases or prepositions?
- Does the language allow free word order?
- etc.

Scanners

- Resolution?

(Overhead scanners are more efficient, flatbed scanners are better quality.)

Topics covered today

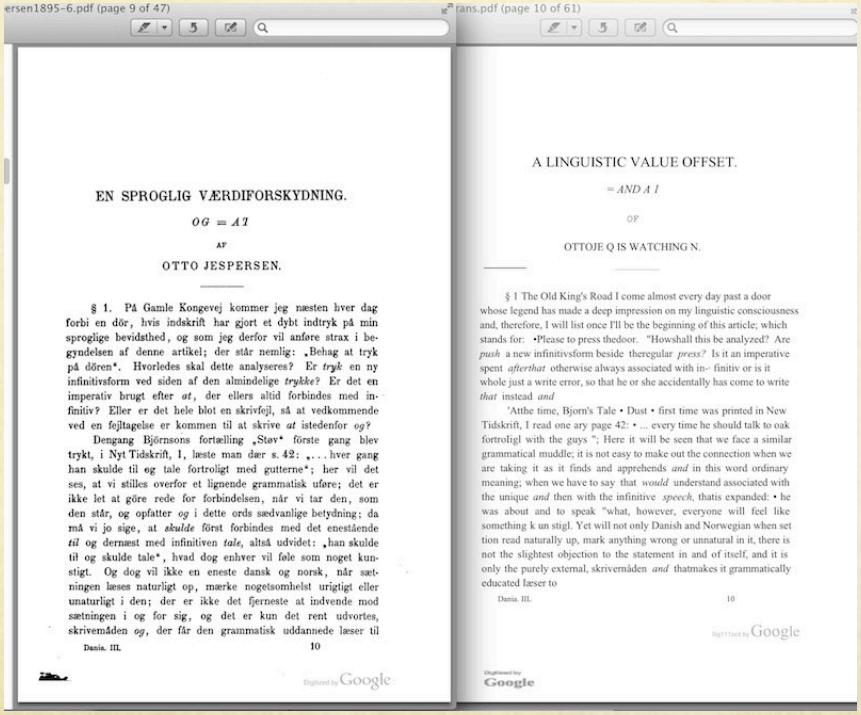
- Optical Character Recognition (OCR)
  - Specifically, Adobe Acrobat and Abbyy Finereader
- Machine Translation (MT)
  - Specifically, Google Translate for demo
- Post-editing
- Hands-on demos
- Discussion

The rest of the presentation was demonstrations of example translation and a hands-on demo for the audience to try.

—*Summarized here.*—

It's possible to translate a document and preserve formatting...

- Using OCR, then Acrobat to convert to Word, then Google Documents and the Translate function...



ersen1895-6.pdf (page 9 of 47)

EN SPROGLIG VÆRDIFORSKYDNING.

OG = AT

AF

OTTO JESPERSEN.

§ 1. På Gamle Kongevej kommer jeg næsten hver dag forbi en dör, hvis indskrift har gjort et dybt indtryk på min sproglige bevidsthed, og som jeg derfor vil anføre strax i begyndelsen af denne artikel; der står nemlig: „Behag at tryk på dören". Hvorledes skal dette analyseres? Er *tryk* en ny infinitivsform ved siden af den almindelige *trykke*? Er det en imperativ brugt efter *at*, der ellers altid forbindes med infinitiv? Eller er det hele blot en skrivfejl, så at vedkommende ved en fejltagelse er kommen til at skrive *at* istedenfor *og*?

Dengang Björnsons fortælling „Stöv" første gang blev trykt, i Nyt Tidskrift, I, læste man dær s. 42: „... hver gang han skulde til og tale fortroligt med gutterne"; her vil det ses, at vi stilles overfor et lignende grammatisk uføre; det er ikke let at göre rede for forbindelsen, når vi tar den, som den står, og opfatter *og* i dette ords sædvanlige betydning; da må vi jo sige, at *skulde* först forbindes med det enestående *til* og dernæst med infinitiven *tale*, altså udvidet: „han skulde til og skulde tale", hvad dog enhver vil føle som noget kunstigt. Og dog vil ikke en eneste dansk og norsk, når sætningen læses naturligt op, mærke nogetsomhelst urigtigt eller unaturligt i den; der er ikke det fjerneste at indvende mod sætningen i og for sig, og det er kun det rent udvortes, skrivemåden *og*, der får den grammatisk uddannede læser til

rans.pdf (page 10 of 61)

A LINGUISTIC VALUE OFFSET.

= AND A I

OF

OTTOJE Q IS WATCHING N.

§ 1 The Old King's Road I come almost every day past a door whose legend has made a deep impression on my linguistic consciousness and, therefore, I will list once I'll be the beginning of this article; which stands for: •Please to press thedoor. "Howshall this be analyzed? Are *push* a new infinitivsform beside theregular *press?* Is it an imperative spent *afterthat* otherwise always associated with in- finitiv or is it whole just a write error, so that he or she accidentally has come to write *that* instead *and*

'Atthe time, Bjorn's Tale • Dust • first time was printed in New Tidskrift, I read one ary page 42: • ... every time he should talk to oak fortrofigl with the guys "; Here it will be seen that we face a similar grammatical muddle; it is not easy to make out the connection when we are taking it as it finds and apprehends *and* in this word ordinary meaning; when we have to say that *would* understand associated with the unique *and* then with the infinitive *speech*, thatis expanded: • he was about and to speak "what, however, everyone will feel like something k un stigl. Yet will not only Danish and Norwegian when set tion read naturally up, mark anything wrong or unnatural in it, there is not the slightest objection to the statement in and of itself, and it is only the purely external, skrivemåden *and* thatmakes it grammatically educated læser to

EN SPROGLIG VÆRDIFORSKYDNING.

OG = AT

AF

OTTO JESPERSEN.

§ 1. På Gamle Kongevej kommer jeg næsten hver dag forbi en dör, hvis indskrift har gjort et dybt indtryk på min sproglige bevidsthed, og som jeg derfor vil anføre strax i begyndelsen af denne artikel; der står nemlig: „Behag at tryk på dören". Hvorledes skal dette analyseres? Er *tryk* en ny infinitivsform ved siden af den almindelige *trykke*? Er det en imperativ brugt efter *at*, der ellers altid forbindes med infinitiv? Eller er det hele blot en skrivfejl, så at vedkommende ved en fejltagelse er kommen til at skrive *at* istedenfor *og*?

Dengang Björnsons fortælling „Støv" förste gang blev trykt, i Nyt Tidskrift, I, læste man dær s. 42: „... hver gang han skulde til og tale fortroligt med gutterne"; her vil det ses, at vi stilles overfor et lignende grammatisk uføre; det er ikke let at göre rede for forbindelsen, når vi tar den, som den står, og opfatter *og* i dette ords sædvanlige betydning; da må vi jo sige, at *skulde* först forbindes med det enestående *til* og dernæst med infinitiven *tale*, altså udvidet: „han skulde til og skulde tale", hvad dog enhver vil føle som noget kunstigt. Og dog vil ikke en eneste dansk og norsk, når sætningen læses naturligt op, mærke nogetsomhelst urigtigt eller unaturligt i den; der er ikke det fjerneste at indvende mod sætningen i og for sig, og det er kun det rent udvortes, skrivemåden *og*, der får den grammatisk uddannede læser til

A LINGUISTIC VALUE OFFSET.

= AND A T

OF

OTTOJE Q IS WATCHING N.

§ 1 The Old King's Road I come almost every day past a door whose legend has made a deep impression on my linguistic consciousness and, therefore, I will list once I'll be the beginning of this article; which stands for: •Please to press thedoor. "Howshall this be analyzed? Are *push* a new infinitivsform beside theregular *press?* Is it an imperative spent *afterthat* otherwise always associated with in- finitiv or is it whole just a write error, so that he or she accidentally has come to write *that* instead *and*

'Atthe time, Bjorn's Tale • Dust • first time was printed in New Tidskrift, I read one ary page 42: • ... every time he should talk to oak fortroligl with the guys "; Here it will be seen that we face a similar grammatical muddle; it is not easy to make out the connection when we are taking it as it finds and apprehends *and* in this word ordinary meaning; when we have to say that *would* understand associated with the unique *and* then with the infinitive *speech,* thatis expanded: • he was about and to speak "what, however, everyone will feel like something k un stigl. Yet will not only Danish and Norwegian when set tion read naturally up, mark anything wrong or unnatural in it, there is not the slightest objection to the statement in and of itself, and it is only the purely external, skrivemåden *and* thatmakes it grammatically educated læser to

Example of a Chinese article

汉语动补结构来源的句法分析*

冯 胜 利

提要 本文在以往研究的基础上提出动补形式来源于秦汉时期的并列式双核心结构，这种双核结构的语法性质是“句法词”而非一般意义上的短语，它是在汉语的韵律规则的促发下动词上移的结果。根据双核的结构要求，[VV]形式如带宾语，其中的自动词必为使动；如果[VV]不带宾语而只有主语，则其中的他动词必为被动。此所以“战败魏师”、“败杀项梁”中的自动词概为使动而“二世杀死”中的他动词必为被动的原因所在。然而，双核结构是不稳定的。双核结构的两解性为其自身由双核向单核的偏移提供了演变的可能性。演变的契机来源于语义的偏重而语义的偏倾则导致了核心的转移。文章指出：双核的偏移采取两种形式，一是由“义素兼并”导致的核心左偏，一是由“语义类差”带来的核心右偏。前者使右边的成分丧失其核心地位而变为补述语，后者则使左边的成分丧失核心地位而变为状语。右边成分核心地位的丧失带来两种必然的结果：首先是使右边他动词的及物性无能为用，于是其指派受事论旨的能力逐渐退化，最终变为自动词；其波所及则使右边自动词的使动用法沦为多余，于是其使动用法也相继消退。两种结果归为一个结构：动补形式应运而生。与此同时，由语义类差带来的核心右偏也日趋活跃，[V_{Adv} V]也便由此而兴，以至为后来大批量的状动复合词创造了新的格式。

关键词 并列结构 核心偏移 动补结构 韵律句法

OCR text (raw):

- 提要本文在以往研究的基础上提出动补形式来源于秦汉时期的并列式 双核心结构,这种双核结构的语法性质是“句法词”而非一般意义上的短 语,它是在汉语的韵律规则的促发下动词上移的结果。根据双核的结构 要求,[乂乂]形式如带宾语,其中的自动词必为使动;如果〔\以〕不带宾语而 只有主语,则其中的他动词必为被动。此所以“战败魏师”、“败杀项梁”中 的自动词概为使动而“二世杀死”中的他动词必为被动的原因所在。然 而,双核结构是不稳定的。双核结构的两解性为其自身由双核向单核的 偏移提供了演变的可能性。演变的契机来源于语义的偏重而语义的偏倾 则导致了核心的转移。文章指出:双核的偏移采取两种形式,一是由“义 素兼并”导致的核心左偏, 一是由“语义类差”带来的核心右偏。前者使右 边的成分丧失其核心地位而变为补述语,后者则使左边的成分丧失核心 地位而变为状语。右边成分核心地位的丧失带来两种必然的结果:首先 是使右边他动词的及物性无能为用,于是其指派受事论旨的能力逐渐退 化,最终变为自动词;其波所及则使右边自动词的使动用法沦为多余,于 是其使动用法也相继消退。两种结果归为一个结构:动补形式应运而生。 与此同时,由语义类差带来的核心右偏也日趋活跃,[乂;^V】也便由此而 兴,以至为后来大批量的状动复合词创造了新的格式。

Google Translate text (raw):

Abstract In this paper, based on previous studies on moving up in the form of parallel twin-core structure from Qin and Han Dynasties, the syntax of the nature of the dual-core architecture is the phrase "syntax Words" rather than a general sense, it is the rhythm of Chinese rule the results of the verb move under the trigger. According to the requirements of the dual-core structure, [qe qe] forms such as belt object, which word will automatically move to make; if [\ to] without object and only subject, the one he will be for a passive verb. This so "defeat Wei Shi" and "defeat Xiang Liang kill" in order to move almost automatic word and "II kill" in his will for the passive verb reason. However, dual-core structure is unstable. Solution of two dual-core structure provides the possibility for their own evolution to shift from single-core dual-core. The opportunity comes from the evolution of semantics and semantic emphasis on the result of the partial core dump transfer. The article points out: Offset duo take two forms, one core left by the "plain meaning of the merger" led to the partial, the first being the "semantic class difference" brings core starboard. The former component to the right so that the loss of its central position and becomes the complement predicate, which makes the loss of the left component of the core position and becomes adverbial. The loss of the right of ingredients to bring the core of the two inevitable result: the first is to make the right side of his impotence transitive verb is used, then its ability to assign patient argument purpose gradual degradation and eventually become automatic word; its waves reach the so that the right of automatic word causative become superfluous, so it makes use of law have also subsided. Classified as a structure of two results: moving up form came into being. At the same time, the semantic class difference brings core right side are becoming increasingly active, [qe; ^ V] they will thus Xing, as well as for the later move large quantities of a compound word like creating a new format.

Result: not perfect, but mostly intelligible.

Published English abstract:

Feng Shengli, A Formal Analysis of the Origin of VR-Constructions in Chinese

Abstract: Based on previous studies, this paper proposes that the so-called VR-constructions in Classical Chinese were originated from double-headed VV structures, motivated by prosodic requirements of the language, resulting in syntactic compounds. It is further argued that the double-headed syntactic compounds are structurally ambiguous, thus, a process of head shift in the VV forms is not only permissible but indeed inevitable in well-defined semantic environments. As a result, when the head was shifted to the left, the right part of the VV compounds became a complement, giving rise to the VR structure in the language, but when the head was shifted to right, the left part of the VV compounds became a modifier, resulting in [VAdv V] compounds in later developments.

Key words: double-headed structure, head-shift, VR-construction, prosodic syntax

Result: the MT version would give us most of the same information if no official English abstract were provided.

Examples from Hands-on demo (Russian):

- Kert (1971) is the only grammar for Kildin Saami, an endangered language spoken in northern Russia.
- It is a thorough and useful reference, but written in Russian.
- Is it accessible through MT?

600dpi color scan (first page):

Предисловие

Настоящая работа представляет собой систематическое описание одного из бесписьменных языков финно-угорской семьи - саамского. В основу описания положен вороньинский говор саамского языка. Село Воронье (по-саамски kårdɨg sijt) расположено в среднем течении р. Вороньей. В силу географического расположения (удаленность от других населенных пунктов), а также по составу населения (имелась только одна русская семья) представители этого села хорошо сохранили особенности своего говора.

Выбор темы исследования объясняется следующими обстоятельствами. В саамском языке агглютинация сочетается с сильно развитой флексией основы. В нем чередуются почти все гласные и согласные звуки. Такая система чередований создала богатый инвентарь фонем. Достаточно сказать, что в саамском языке насчитывается более ста фонем. Таким образом, строй саамского языка дает интересный материал для постановки и решения таких проблем общего языкознания, как проблема типологической классификации языков(соотношение флексии и агглютинации), проблема интерпретации фонем (дистрибуция и функционирование),и ряда других.И, конечно,факты саамского языка дают чрезвычайно ценный материал для сравнительно-исторического изучения финно-угорских языков.

OCR text (raw):

Настоящая рабой представляет собой систематическое описание одного из бесписьменных языков финно-угорской семьи - саамского. В основу описания положен вороньинский говор саамского языка. Се- ло Воронье (по-саамски к!га*в в*?*) расположено в среднем тече- нии р. Вороньей. В силу географического расположения (удаленность от других населенных пунктов), а также по составу населения (име- лась только одна русская семья) представители этого села хорошо сохранили особенности своего говора.

Выбор темы исследования объясняется следующими обстоятельст- вами. В саамском языке агглютинация сочетается с сильно развитой флексией основы. В нем чередуются почти все гласные и согласные звуки. Такая система чередований создала богатый инвентарь фонем. Достаточно сказать, что в саамском языке насчитывается более ста фонем. Таким образом, строй саамского языка дает интересный мате- риал для постановки и решения таких проблем общего языкознания, как проблема типологической классификации языков(соотношение фле- ксии и агглютинации), проблема интерпретации фонем (дистрибуция*

и функционирование)^ ряда других.И, конечно,факты саамского язы-

ка дают чрезвычайно ценный материал для сравнительно-историческо- го изучения финно-угорских языков.

Minor manual re-formatting:

Настоящая рабой представляет собой систематическое описание одного из бесписьменных языков финно-угорской семьи - саамского. В основу описания положен вороньинский говор саамского языка. Се- ло Воронье (по-саамски KARDEG SIJT) расположено в среднем тече- нии р. Вороньей. В силу географического расположения (удаленность от других населенных пунктов), а также по составу населения (име- лась только одна русская семья) представители этого села хорошо сохранили особенности своего говора.

Выбор темы исследования объясняется следующими обстоятельст- вами. В саамском языке агглютинация сочетается с сильно развитой флексией основы. В нем чередуются почти все гласные и согласные звуки. Такая система чередований создала богатый инвентарь фонем. Достаточно сказать, что в саамском языке насчитывается более ста фонем. Таким образом, строй саамского языка дает интересный мате- риал для постановки и решения таких проблем общего языкознания, как проблема типологической классификации языков(соотношение фле- ксии и агглютинации), проблема интерпретации фонем (дистрибуция* и функционирование)^ ряда других.И, конечно,факты саамского языка дают чрезвычайно ценный материал для сравнительно-исторического изучения финно-угорских языков.

You can try it yourself now!
(Just copy and past the text.)

Google Translate (raw):

This slave is a systematic description of one of the unwritten languages of the Finno-Ugric family - Sami. The basis of the description put voroninsky speaking the Sami language. Se lo Raven (in Sámi KARDEG SIJT) located in the middle Techa River Scientific Research Institute. Ravens. Due to the geographical location (distance from other settlements), as well as the composition of the population (IME familiarize only one Russian family) members of this village is well preserved features of the contract.

The choice of research topic due to the following you through circumstances. The Sami agglutination combined with highly developed inflexion base. It alternated nearly all the vowels and consonants. This system has created a wealth of alternations inventory of phonemes. Suffice it to say that in the Sami language, there are over a hundred of phonemes. Thus, the structure of the Sami language provides an interesting mate rial for formulating and solving such problems of general linguistics, as the problem of typological classification of languages (the ratio of fle FIAC and agglutination), the problem of interpretation of phonemes (distribution and function *) ^ number drugih.I, of course, the facts Sami language provide extremely valuable material for comparative-historical study of the Finno-Ugric languages.

Some minor errors, but generally intelligible.

Post-editing:

"This **slave** is a systematic description of one of the unwritten languages of the Finno-Ugric family – Sami."

- Just one example for now, but consider the bold word 'slave' here. It makes no sense in context, so any English speaker (especially a linguist with some basic expectations for what will be written in a grammar) can determine that it isn't the right translation.

Настоящая **рабой** представляет собой систематическое описание одного из бесписьменных языков финно-угорской семьи - саамского.

- Using Google Translate, if we work on just this sentence, we can highlight the word "**slave**" in the translated text and find that it corresponds to "**рабой**" in the Russian original.

Post-editing:

- In most cases, highlighting this word would reveal better choices in the context. This works for other portions of the text. But in this case, that method doesn't work for this word.
- If we try putting the word "**рабой**" into Wiktionary, no results are located. But if we try to guess the stem of the word by removing the last letter, we get the suggestion "**работа**". We can translate that word with Google Translate now, or use a dictionary.
- This word turns out to mean "work, job, labor, service". This makes a lot more sense in context. It's not exactly clear what was going on in the original sentence, but we can be fairly confident in the following translation:
- "This **work** is a systematic description of one of the unwritten languages of the Finno-Ugric family – Sami."

Post-editing:

- Most post-editing corrections are not that difficult. Some may be, and some may be impossible without the help of a native speaker. (Or even a native speaker domain expert!)
- For the most part, we can translate the text with reasonable quality.
- Some changes will take a little more time and patience. The limitations with this method are speed, not knowledge of Russian.
- Therefore, use it to identify interesting content, not necessarily to translate full works at high quality. Checking any final translations with a native speaker is a good idea, but you can save their time and your money if you do most of the intermediate work yourself. It's always better to ask a translator to translate one paragraph than a whole article.

Materials (try it yourself)

- The following pages are an OCR-image and OCR text-only version of two pages from the original scan, using Abbyy Finereader.
- You can try to translate the remaining content yourself by copying and pasting the text into Google Translate and post-editing the results.
- Hints:
  - The remaining paragraphs of this introduction are not too difficult.
  - Remember the context: this is the introduction to a descriptive grammar, so look for clues about its location, speakers and grammatical properties.

Предисловие

Настоящая работа представляет собой систематическое описание одного из бесписьменных языков финно-угорской семьи - саамского. В основу описания положен вороньинский говор саамского языка. Село Воронье (по-саамски kårdɨg sijt) расположено в среднем течении р. Вороньей. В силу географического расположения (удаленность от других населенных пунктов), а также по составу населения (имелась только одна русская семья) представители этого села хорошо сохранили особенности своего говора.

Выбор темы исследования объясняется следующими обстоятельствами. В саамском языке агглютинация сочетается с сильно развитой флексией основы. В нем чередуются почти все гласные и согласные звуки. Такая система чередований создала богатый инвентарь фонем. Достаточно сказать, что в саамском языке насчитывается более ста фонем. Таким образом, строй саамского языка дает интересный материал для постановки и решения таких проблем общего языкознания, как проблема типологической классификации языков(соотношение флексии и агглютинации), проблема интерпретации фонем (дистрибуция и функционирование),и ряда других.И, конечно,факты саамского языка дают чрезвычайно ценный материал для сравнительно-исторического изучения финно-угорских языков.

В процессе исторического развития в словарном составе саамского языка отложились различные пласты лексики, в том числе и из языков других систем (литво-латышские, германские, славянские).В то же время саамский язык характеризуется значительным слоем субстратной лексики, которая не имеет соответствий ни в одном из современных живых языков. Исследование грамматического строя и лексического состава саамского языка поможет в известной мере проследить его историю,а вместе с тем историю носителя этого языка - саамского народа.А история саамов, по нашему глубокому убеждению, является ключом к разгадке истории европейского Севера.

Что касается исходных принципов анализа саамского языка, то мы следовали грамматической традиции московской лингвистической школы , основателем которой был академик Ф.Ф.Фортунатов. При описании фактов языка мы старались идти от формы к значению , хотя это и не всегда удавалось. Раздел морфологии дается нами в традиционном плане описания частей речи,несмотря на убеждение в непоследовательности применения принципов классификации фактов языка по частям речи.При изучении синтаксических явлений необходимо было предварительно рассмотреть вопрос о соотношении языка и мышле-

ния. Мы вынуждены были отказаться от существующего подхода к этой проблеме, - когда вопрос об отношении языка и мышления сводится в основном к вопросу о соотношении суждения как формы мысли и предложения как языковой единицы. Поскольку структура суждения несопоставима со структурой предложения, отпала необходимость приравнивания членов суждения к главным членам предложения. Предложение является языковой категорией и должно определяться по языковым признакам.

Работа состоит из разделов "Фонетика","Морфология" и"Синтаксис". Исследование способов развития лексического состава путем словообразования и заимствований будет дано особо.

Работа написана на основе материалов , собранных автором во время экспедиций на Кольский полуостров в 1954 -1965 гг., а также материалов по кильдинскому диалекту, опубликованных Т.Итконеном в сборнике сказок "Koltan- ja Kuolan lappalaisia satuja" и в словаре А.Генетца "Kuollan lapin murteiden sanakirja ynnä kielennäytteitä". Использована также учебная литература, изданная в период создания саамской письменности.

Автор искренне благодарен проф. М.И.Матусевич, доктору филологических наук В.З.Панфилову, а также сотрудникам сектора языкознания Института языка, литературы и истории Карельского филиала АН СССР и всем лицам, взявшим на себя труд по прочтению настоящей рукописи, советы и пожелания которых способствовали улучшению работы.

Особенно автор благодарит саамов Т.В.Матрехину, А.А.Антонову,Г.А.Шаршину,В.Н.Железнякову,С.Г.Юрьева,Ф.Н.Яковлева, В.Г.Кобелева и др., оказавших ему большую помощь в сборе и систематизации материалов по саамскому языку.

П р е д и с л о в и е

Настоящая рабой представляет собой систематическое описание одного из бесписьменных языков финно-угорской семьи - саамского. В основу описания положен вороньинский говор саамского языка. Село Воронье (по-саамски к!га*в в*?*) расположено в среднем течении р. Вороньей. В силу географического расположения (удаленность от других населенных пунктов), а также по составу населения (имелась только одна русская семья) представители этого села хорошо сохранили особенности своего говора.

Выбор темы исследования объясняется следующими обстоятельствами. В саамском языке агглютинация сочетается с сильно развитой флексией основы. В нем чередуются почти все гласные и согласные звуки. Такая система чередований создала богатый инвентарь фонем. Достаточно сказать, что в саамском языке насчитывается более ста фонем. Таким образом, строй саамского языка дает интересный материал для постановки и решения таких проблем общего языкознания, как проблема типологической классификации языков(соотношение флексии и агглютинации), проблема интерпретации фонем (дистрибуция* и функционирование)^ ряда других.И, конечно,факты саамского языка дают чрезвычайно ценный материал для сравнительно-исторического изучения финно-угорских языков.

В процессе исторического развития в словарном составе саамского языка отложились различные пласты лексики, в том числе и из языков других систем (литво-латышские, германские, славянские).В то же время саамский язык характеризуется значительным слоем субстратной лексики, которая не имеет соответствий ни в одном из современных живых языков. Исследование грамматического строя и лексического состава саамского языка поможет в известной мере проследить его историю,а вместе с тем историю носителя этого языка - саамского народа.А история саамов, по нашему глубокому убеждению, является ключом к разгадке истории европейского Севера.

Что касается исходных принципов анализа саамского языка, то мы следовали грамматической традиции московской лингвистической школы , основателем которой был академик Ф.Ф.Фортунатов. При описании фактов языка мы старались идти от формы к значению , хотя это и не всегда удавалось. Раздел морфологии дается нами в традиционном плане описания частей речи,несмотря на убеждение в непоследовательности применения принципов классификации фактов языка по частям речи.При изучении синтаксических явлений необходимо было предварительно рассмотреть вопрос о соотношении языка и мышле-

ния. Мы вынуждены были отказаться от существующего подхода к этой проблеме, - когда вопрос об отношении языка и мышления сводится в основном к вопросу о соотношении суждения как формы мысли и предложения как языковой единицы. Поскольку структура суждения несопоставима со структурой предложения, отпала необходимость приравнивания членов суждения к главным членам предложения. Предложение является языковой категорией и должно определяться по языковым признакам.

Работа состоит из разделов "Фонетика","Морфология" ^'Синтаксис". Исследование способов развития лексического состава путем словообразования и заимствований будет дано особо.

Работа написана на основе материалов , собранных автором во время экспедиций на Кольский полуостров в 1954 -1965 гг., а также материалов по кидьдивскому диалекту, опубликованных Т.Итконеном в Сборнике сказок "КоИап- за Кио1ап 1арра1а1з1а зализа" И В словаре А.Генетца "КиоПап 1ар1п тшгЬеЛАеп запак3.гза уппа к1в1еппауь-ье1-ьа". Использована также учебная литература, изданная в период создания саамской письменности.

Автор искренне благодарен проф. М.И.Матусевич, доктору филологических наук В.З.Панфилову, а также сотрудникам сектора языкознания Института языка, литературы и истории Карельского филиала АН СССР и всем лицам, взявшим на себя труд по прочтению настоящей рукописи, советы и пожелания которых способствовали улучшению работы.

Особенно автор благодарит саамов Т.В.Матрехину, А.А.Антонову,Г.А.Шаршину,В.Н.1елезнякову,С.Г.Юрьева,Ф.Н.Яковлева, В.Г.Кобелева и др., оказавших ему большую помощь в сборе и систематизации материалов по саамскому языку.

References

- Lane Schwartz, Timothy Anderson, Jeremy Gwinnup, and Katherine M. Young. 2014. Machine translation and monolingual postediting: The AFRL WMT-14 system. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 186–194, Baltimore, Maryland, June. Association for Computational Linguistics.

- Lane Schwartz. 2014. Monolingual post-editing by a domain expert is highly effective for translation triage. In *Proceedings of the The Third Workshop on Post-editing Technology and Practice*, Vancouver, Canada, October.

- Jespersen, O. (1895). En sproglig værdiforskydning: og = at. *Dania: Tidsskrift for Folkemål Og Folkeminder, 3*, 145–183.

- Feng, S. [冯胜利]. (2002). 汉语动补结构来源的句法分析. In 北京大学汉语语言学研究中心《语言学论丛》编委会, 语言学论丛. 第二十六辑 第二十六辑 (pp. 178–208). Beijing: 商务印书馆.

- Kert, G. M. (1971). *Саамский язык (Кильдинский Диалект): фонетика, морфология, синтаксис* [Saami language (Kildin dialect): phonetics, morphology, syntax]. Leningrad: Nauka.

Questions?

- Not everything from the talk could be preserved in this PDF document available on the web.

- If you have questions or want more information about using MT for these purposes, feel free to contact me.

djross3@illinois.edu