Language typology and language description: General and Uralic perspectives
Saaren kartano / Kone Foundation, 12 June 2015
Matti Miestamo, University of Helsinki

The world’s languages

Source: The Ethnologue 16th edn.
<http://archive.ethnologue.com/16/show_map.asp?name=World&seq=10>

Geographical distribution

<table>
<thead>
<tr>
<th>Region</th>
<th>Living languages</th>
<th>Number of speakers</th>
<th>Percent</th>
<th>Median</th>
<th>Typical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>2,135</td>
<td>95.1</td>
<td>12.0</td>
<td>361,316</td>
<td>27.52K</td>
</tr>
<tr>
<td>Americas</td>
<td>3,094</td>
<td>90.0</td>
<td>0.9</td>
<td>404,426</td>
<td>1,142K</td>
</tr>
<tr>
<td>Asia</td>
<td>2,361</td>
<td>51.4</td>
<td>48.1</td>
<td>1,642,963</td>
<td>12,225K</td>
</tr>
<tr>
<td>Europe</td>
<td>290</td>
<td>4.0</td>
<td>26.0</td>
<td>9,727,252</td>
<td>28,905K</td>
</tr>
<tr>
<td>Pacific</td>
<td>1,213</td>
<td>10.5</td>
<td>0.1</td>
<td>1,146</td>
<td>904</td>
</tr>
<tr>
<td>World</td>
<td>7,602</td>
<td>100.0</td>
<td>100.0</td>
<td>868,454</td>
<td>7,805K</td>
</tr>
</tbody>
</table>

(Data from Lewis et al. 2015)

Numbers of languages and speakers

<table>
<thead>
<tr>
<th>Languages and speakers</th>
<th>speakers per lg</th>
<th>languages</th>
<th>speakers total</th>
</tr>
</thead>
<tbody>
<tr>
<td>over 100 million</td>
<td>8 (0.1 %)</td>
<td>2.5 billion (40.2 %)</td>
<td></td>
</tr>
<tr>
<td>over 1 million</td>
<td>394 (5.5 %)</td>
<td>5.9 billion (94.2 %)</td>
<td></td>
</tr>
<tr>
<td>less than 10,000</td>
<td>3,731 (52.5 %)</td>
<td>8.1 million (0.13 %)</td>
<td></td>
</tr>
</tbody>
</table>

(Data from Lewis et al. 2015)
Numbers of languages and speakers

<table>
<thead>
<tr>
<th>Numbers of languages and speakers</th>
<th>Count</th>
<th>Percent</th>
<th>Cumulative</th>
<th>Total</th>
<th>Percent</th>
<th>Cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-100</td>
<td>5</td>
<td>0.1%</td>
<td>5.1%</td>
<td>2,392,420,570</td>
<td>0.1%</td>
<td>5.1%</td>
</tr>
<tr>
<td>100-1,000</td>
<td>324</td>
<td>4.5%</td>
<td>5.1%</td>
<td>2,402,709,970</td>
<td>4.5%</td>
<td>9.6%</td>
</tr>
<tr>
<td>1,000-10,000</td>
<td>1,002</td>
<td>20.7%</td>
<td>25.7%</td>
<td>35,303,485</td>
<td>18.0%</td>
<td>43.7%</td>
</tr>
<tr>
<td>10,000-100,000</td>
<td>10,000</td>
<td>20.7%</td>
<td>46.4%</td>
<td>81,902,737</td>
<td>48.5%</td>
<td>95.1%</td>
</tr>
<tr>
<td>100,000-1000,000</td>
<td>100,000</td>
<td>0.0%</td>
<td>0.0%</td>
<td>5,623,323</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>1,000,000-5,000,000</td>
<td>7,652</td>
<td>0.1%</td>
<td>0.1%</td>
<td>96,368,967</td>
<td>0.0%</td>
<td>0.1%</td>
</tr>
<tr>
<td>5,000,000-100,000</td>
<td>5,000</td>
<td>0.0%</td>
<td>0.0%</td>
<td>91,622,095</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>100,000,000-500,000</td>
<td>100,000</td>
<td>0.0%</td>
<td>0.0%</td>
<td>91,622,095</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>500,000,000-1,000,000</td>
<td>500,000</td>
<td>0.0%</td>
<td>0.0%</td>
<td>100,000,000</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

(Lewis et al. 2015)

Language vitality and endangered languages

**Language endangerment**

- "[T]he coming century will see either the death or the doom of 90% of mankind’s languages." (Krauss 1992: 7)
- "It is estimated that, if nothing is done, half of 6000 plus languages spoken today will disappear by the end of this century." (UNESCO)
- The situation involves a multitude of historical, political, social factors and problems.

**Language vitality (UNESCO)**

- 2471 languages classified as other than safe by UNESCO.

**Language vitality (Ethnologue)**

- 2471 languages classified as other than safe by UNESCO.
Language vitality

Table 4. Distribution of world languages by vitality status

<table>
<thead>
<tr>
<th>Region</th>
<th>Total</th>
<th>Low</th>
<th>Medium-Low</th>
<th>Medium-High</th>
<th>High</th>
<th>Very High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>1,171,592</td>
<td>1,151,674</td>
<td>11,941</td>
<td>648</td>
<td>40</td>
<td>1</td>
</tr>
<tr>
<td>Americas</td>
<td>910,100</td>
<td>691,100</td>
<td>219,000</td>
<td>67,000</td>
<td>3,000</td>
<td>250</td>
</tr>
<tr>
<td>Asia</td>
<td>3,278,300</td>
<td>3,207,600</td>
<td>70,700</td>
<td>51,000</td>
<td>10,000</td>
<td>1,500</td>
</tr>
</tbody>
</table>

(Preliminary estimates according to Siegl & Grünthal, 2012)

<http://listserv.linguistlist.org/pipermail/ura-list/2012-October/001269.html>

Language vitality by continent

Uralic languages

Figure 1. Map of Uralic languages (design by György Liszka)

(Preliminary estimates according to Siegl & Grünthal, 2012)

<http://listserv.linguistlist.org/pipermail/ura-list/2012-October/001269.html>

Uralic languages

[Tables containing speaker data omitted; can be found via the link below]

(Preliminary estimates according to Siegl & Grünthal, 2012)

<http://listserv.linguistlist.org/pipermail/ura-list/2012-October/001269.html>

Uralic languages

[Tables containing speaker data omitted; can be found via the link below]

(Preliminary estimates according to Siegl & Grünthal, 2012)

<http://listserv.linguistlist.org/pipermail/ura-list/2012-October/001269.html>
Endangered Uralic languages (UNESCO)

- Vulnerable
  - North Saami
  - Karelian (Karelia), Karelian (Tver), Olonetsian, Våro-Sesto
  - Erzya, Moksha
  - Eastern Mari
  - Komi, Permyak, Udmurt
  - Eastern Khanty, Northern Khanty
  - Tundra Nenets
- Definitely endangered
  - Ingrı́, Karelian (Tikhvin), Lude, Veps
  - Inari Saami, Skolt Saami, Lule Saami, South Saami, Kidinn Saami, Western Mari
  - Yaova Komi
  - Csángo Hungarian
  - Northern Mansi
  - Njarasam, Forest Nenets, Northern Selkup
- Severely endangered
  - Livonian, Vote
  - Pite Saami, Ume Saami, Ter Saami
  - Eastern Mansi
  - Central Selkup, Southern Selkup, Forest Enets, Tundra Enets

Endangered Uralic languages (UNESCO)

- Critically endangered
  - Livonian, Vote
  - Pite Saami, Ume Saami, Ter Saami
  - Eastern Mansi
  - Central Selkup, Southern Selkup, Forest Enets, Tundra Enets
- Extinct
  - Karelian (Valday)
  - Akkala Saami
  - Southern Khanty, Southern Mansi, Western Mansi
  - Kamas

⇒ None in the vulnerable category
⇒ Only Finnish, Estonian and Hungarian would be classified as safe.

The world’s languages – state of documentation

Why documentation/description?

- Given the endangered status of a large number of the world’s languages, documentation of endangered languages is an urgent priority in linguistics
  - to enable revitalization
  - to save the empirical basis of language typology and general linguistics
  - ... and ultimately our possibilities to properly understand the nature of human language (and the human mind / human behaviour).
  - to save the non-linguistic knowledge carried by dying languages.
- Serves the community
  - revitalization
  - education
  - language planning
  - identity and attitudes.
- Scientific value
  - knowledge about human language (and more generally about the human mind) presupposes information on individual languages
  - language typology and descriptive linguistics need each other
  - typology => theoretical basis for description
  - description => empirical basis for typology

Documentation vs. description

Documentation of a language is an activity (and, derivatively, its result) that gathers, processes and exhibits a sample of data of the language that is representative of its linguistic structure and gives a fair impression of how and for what purposes the language is used. Its aim is to represent the language for those who do not have access to the language itself. (Lehmann 2001: 83)
Typology and description

**Documentation vs. description**

**Description** of a language is an activity (and, derivatively, its result) that formulates, in the most general way possible, the patterns underlying the linguistic data. Its aim is to make the user of the description understand the way the language works. (Lehmann 2001: 83)

- Documentation in a broad sense includes both aspects.

**Documentation of a language**

- Lectodoc / doculect
- A basic description of a language includes a grammar, a text collection and a dictionary.
- The world’s languages have been documented to different extents (Hammarström 2007):
  - Wordlist: a wordlist of some length.
  - Phonology: a wordlist with a phonological statement, typically 20 pages.
  - Sketch: wordlist, phonology plus major aspects of morphology and syntax, typically 50 pages.
  - Short grammar: Some treatment of all significant aspects of the language, typically 100 pages.
  - (Full-length) grammar: In-depth description of all significant aspects of the language, typically 300 pages.

**Document types (Glottolog)**

- **Bibliographical**
  - bibliographical information (i.e., the language is featured in a bibliography)
- **Comparative**
  - the language is featured in a comparative study
- **Dialectology**
  - containing dialectological information, e.g., the intelligibility between different dialects, the distribution of certain isoglosses within a language
- **Dictionary**
  - ~ 75 pages and beyond
- **Ethnographic**
  - ethnographic information (whether extensive or brief)
- **Grammar**
  - an extensive description of most elements of the grammar ~ 150 pages and beyond
- **Grammar Sketch**
  - a less extensive description of many elements of the grammar ~ 50 pages
- **Minimal**
  - some small amount of lexical or grammatical data but not sufficient for a full wordlist or a substantial account of some grammatical feature

- **New Testament**
  - a new testament translation
- **Overview**
  - the language is featured in a handbook/overview publication
- **Phonology**
  - phonological description
- **Sociolinguistic**
  - sociolinguistic information (where spoken, by how many etc)
- **Specific Feature**
  - description of some element of grammar (i.e., noun class system, verb morphology etc)
- **Text**
  - some amount of unanalyzed text data ~ 10 pages and beyond
- **Wordlist**
  - wordlist ~ a couple of hundred words

[<http://glottolog.org>]

**Document types (Glottolog)**

- **New Testament**
  - a new testament translation
- **Overview**
  - the language is featured in a handbook/overview publication
- **Phonology**
  - phonological description
- **Sociolinguistic**
  - sociolinguistic information (where spoken, by how many etc)
- **Specific Feature**
  - description of some element of grammar (i.e., noun class system, verb morphology etc)
- **Text**
  - some amount of unanalyzed text data ~ 10 pages and beyond
- **Wordlist**
  - wordlist ~ a couple of hundred words

[<http://glottolog.org>]

**Documentation of the world’s languages**

- Number of languages documented at different levels (published work):
  - Wordlist 4,729
  - Sketch 3,337
  - (Short or full) grammar 2,215 (Hammarström 2007)
- Number of languages with bible translations (as of 31.12.2008):
  - Portions 843
  - Testaments 1,185
  - Bibles 451
  - Total 2,479 (data from UBS)
- Glottolog / LangDoc [http://glottolog.org/]

**Least documented families**

- According to Hammarström (2012), 27 language families satisfy the following criteria
  - The language family is known through at least a wordlist (i.e., this excludes languages known to exist, but for which there are no data, such as the languages of isolados).
  - The language family, at the present state of knowledge, is not demonstrably related to any other known family.
  - There are no viable grounds for concluding that the language is extinct, i.e., that it does not have fluent speakers.
  - All languages of the family are poorly documented, in the sense that there is less documentation than a rudimentary grammar sketch, and there is no ongoing documentation effort for any of them.
- Mostly New Guinea, South America.
- These are entire families (mostly isolates though); underdocumented languages abound in other families as well.
Reasons for the poor state of documentation

- Interest in dominant languages.
- Prestige of theoretical work, low esteem of data collection.
- Lack of funds, high cost of fieldwork.
- Difficulty of fieldwork.
- Is the bad state of description a myth?
  – Nordhoff & Hammarström 2012: "grey literature"

Recent developments

- During the last 20 years, the state of documentation of the world’s languages has improved significantly.
  – Dobes, Rausing, Kone Foundation.
  – Glottolog/Langdoc.
  – Still, only a small portion of the world’s linguistic diversity has been documented.
- The availability of sources has improved
  – electronic archiving and dissemination
  – communication between experts is easier than before.
- New types of documentation materials/methods
  – Digital resources
  – Software
  – Data collection techniques
  – Mobile applications etc.

Recent developments

- During the last 20 years, the state of documentation of the world’s languages has improved significantly.
  – Dobes, Rausing, Kone Foundation.
  – Glottolog/Langdoc.
  – Still, only a small portion of the world’s linguistic diversity has been documented.
- The availability of sources has improved
  – electronic archiving and dissemination
  – communication between experts is easier than before.
- New types of documentation materials/methods
  – Digital resources
  – Software
  – Data collection techniques
  – Mobile applications etc.

Documentation of Uralic languages

- Long tradition in Finno-Ugric studies starting from the work of Sjögren and Castrén in the first half of the 19th century.
- One of the best studied families in the world.
- During Soviet times, fieldwork was not possible for outsiders.
- Historical-comparative focus
- Descriptive/documentary materials typically:
  – text collections
  – dictionaries
  – chrestomathies
  – also grammars
- But modern grammars, written in a typological-functional perspective, are largely lacking.
- "Grey literature"
Uralic languages in Glottolog
- 44 Uralic languages
- 1540 entries in the database
- 122 grammars
- 140 grammar sketches
- Note: Most of the entries are not classified to any category and some of them are classified automatically (=> errors).

Uralic languages with (a) grammar(s) listed in Glottolog
- Estonian, Karelian, Liv, Finnish, Tornedalen Finnish, Veps, Votic
- Skolt Saami, Akkala Saami, Kildin Saami, Ter Saami, Lule Saami, Northern Saami, Pite Saami, South Saami
- Erzya, Moksha
- Eastern Mari, Western Mari
- Komi-Permyak, Komi Zyrian, Udmurt
- Far Eastern Khanty, Khanty, Mansi, Hungarian
- Tundra Nenets, Selkup, Nganasan
  [NB: Forest Enets missing here]

Uralic languages with no grammar listed in Glottolog
- Grammar sketch exists
  - Ingrian, Livvi, Ludian, Kven
  - Ume Saami
  - Surgut Khanty, Southern Khanty
  - Forest Enets [sic!], Forest Nenets, Kamas, Mator
- No grammar sketch
  - Inari Saami, Kemi Saami
  - Tundra Enets, Samoyed Koibal

Uralic languages with no grammar listed in Glottolog
- Grammar sketch exists
  - Ingrian, Livvi, Ludian, Kven
  - Ume Saami
  - Surgut Khanty, Southern Khanty
  - Forest Enets [sic!], Forest Nenets, Kamas, Mator
- No grammar sketch
  - Inari Saami, Kemi Saami
  - Tundra Enets, Samoyed Koibal

Some examples of resources in the category "grammars" in Glottolog
- Livonian
- Erzya
  - Mosun, Michael Vasilievitch and Balakir. 1983. Ersämordvan oppiäry. (Apunayrjaa suomea-ugrisa-samojedaa kielten opinnen varten.) Helsinki: Suomen-ugrisa-samojedaa suoera. 1+188pp. (Audio-Cassette has been transcribed into Audio-CD - please ask library staff.)
- Tundra Nenets

What to make of this?
- A proper literature research would be needed, but the data drawn from Glottolog is an approximation of the situation, and it gives an idea of which languages are most in need of documentation.
- And it also shows what the situation looks like to a non-Uralicist linguist.
- Apart from Estonian, Finnish and Hungarian, few Uralic languages possess a basic description according to modern standards.
- Interesting in view of the long research history of the Uralic family.

Language documentation and language typology
Typology and description

- Symbiotic relationship
  - typology feeds description with theory and ideas
  - descriptive linguistics provides the empirical foundations for typology.
  - This is threatened by the extinction of undescribed languages!

  In other words:
  - Language typology, and thus also general linguistics, is dependent on data provided by researchers focusing on particular languages.
  - who then, in turn, benefit from typological knowledge when trying to understand the phenomena they encounter in their own languages of study.
- They share the common goal of unraveling, describing and understanding the world’s linguistic diversity

  => Diversity Linguistics

Typology & Uralic: Examples of topics

- Negation
- Evidentiality
  - Typology: Chafe & Nichols (1986), De Haan (1997), Aikhenvald & Dixon (2003), Aikhenvald (2004); Kittilä’s project at UH
  - Uralic: ongoing work, e.g., in Kittilä’s project.

Topics in the questionnaire

- Clausal negation
  - standard negation
  - negation in non-declaratives
  - negation of stative predications
  - negation in dependent clauses.
- Non-clausal negation
  - negative replies
  - negation of indefinite pronouns
  - negation of case, derivation and adpositions
- Further aspects of negation
  - scope of negation
  - negative polarity
  - negation and case marking
  - reinforcing negation
  - negation and complex sentences.

What is a "good", "modern" description?

- Typologically informed
  - background for understanding language-particular phenomena, in particular:
    - typology helps to ask relevant questions: fill gaps, see connections
    - provides a metalanguage for description.
    - Payne (1997), Shopen (2007) etc.
  - Functionally oriented
    - organized according to function rather than form
    - who is the grammar written for?
    - cf the contents of Epps: Grammar of Hup (2008)

- Accessible
  - Basic Linguistic Theory (Dixon 1997; Dryer 2006)
  - Framework-free (cf. Haspelmath 2010a)
  - Categorial particularism (cf. Dryer 1997; Haspelmath 2010b; Miestamo 2013)
    - Descriptive categories are language specific
    - focus on properties rather than terminology!
    - Fieldwork as unlearning (Gil 2001)

  => Training is needed!
  A summer school?
Back to the Uralic perspective

- What is the relation to existing tradition?
- Cf. 4-year documentary projects by individuals in previously unexplored territories.
- Is this feasible in Uralic studies?
- Or do we need more extensive projects?
- Unanalysed archive materials into use!

References


Thank you!