Language death in the digital age

András Kornai
HAS Computer Science Research Institute
and
Department of Computer Science, Boston University

@META-FORUM

June 20 2012

Defining language death

- Loss of function (other languages take over entire functional areas)
- Loss of prestige
- Loss of competence (emergence of semi-speakers) Menomini Bloomfield 1927, Gaelic Dorian 1981, Dyrbal Schmidt 1985

In the digital age:

- Loss of function performed digitally that is, increasingly, every function, from day to day communication (texting, email, ...) to commerce, official business, ...
- Loss of prestige If it's not on the web, it doesn't exist
- Loss of competence Can you raise a digital native in your language?

Defining language death

- Loss of function (other languages take over entire functional areas)
- Loss of prestige
- Loss of competence (emergence of semi-speakers) Menomini Bloomfield 1927, Gaelic Dorian 1981, Dyrbal Schmidt 1985

In the digital age:

- Loss of function performed digitally that is, increasingly, every function, from day to day communication (texting, email, ...) to commerce, official business, ...
- Loss of prestige If it's not on the web, it doesn't exist
- Loss of competence Can you raise a digital native in your language?

Defining language death

- Loss of function (other languages take over entire functional areas)
- Loss of prestige
- Loss of competence (emergence of semi-speakers) Menomini Bloomfield 1927, Gaelic Dorian 1981, Dyrbal Schmidt 1985

In the digital age:

- Loss of function performed digitally that is, increasingly, every function, from day to day communication (texting, email, ...) to commerce, official business, ...
- Loss of prestige If it's not on the web, it doesn't exist
- Loss of competence Can you raise a digital native in your language?

Defining language death

- Loss of function (other languages take over entire functional areas)
- Loss of prestige
- Loss of competence (emergence of semi-speakers) Menomini Bloomfield 1927, Gaelic Dorian 1981, Dyrbal Schmidt 1985

In the digital age:

- Loss of function performed digitally that is, increasingly, every function, from day to day communication (texting, email, ...) to commerce, official business, ...
- Loss of prestige If it's not on the web, it doesn't exist
- Loss of competence Can you raise a digital native in your language?

Defining language death

- Loss of function (other languages take over entire functional areas)
- Loss of prestige
- Loss of competence (emergence of semi-speakers) Menomini Bloomfield 1927, Gaelic Dorian 1981, Dyrbal Schmidt 1985

In the digital age:

- Loss of function performed digitally that is, increasingly, every function, from day to day communication (texting, email, ...) to commerce, official business, ...
- Loss of prestige If it's not on the web, it doesn't exist
- Loss of competence Can you raise a digital native in your language?

Defining language death

- Loss of function (other languages take over entire functional areas)
- Loss of prestige
- Loss of competence (emergence of semi-speakers) Menomini Bloomfield 1927, Gaelic Dorian 1981, Dyrbal Schmidt 1985

In the digital age:

- Loss of function performed digitally that is, increasingly, every function, from day to day communication (texting, email, ...) to commerce, official business, ...
- Loss of prestige If it's not on the web, it doesn't exist
- Loss of competence Can you raise a digital native in your language?

Defining language death

- Loss of function (other languages take over entire functional areas)
- Loss of prestige
- Loss of competence (emergence of semi-speakers) Menomini Bloomfield 1927, Gaelic Dorian 1981, Dyrbal Schmidt 1985

In the digital age:

- Loss of function performed digitally that is, increasingly, every function, from day to day communication (texting, email, ...) to commerce, official business, ...
- Loss of prestige If it's not on the web, it doesn't exist
- Loss of competence Can you raise a digital native in your language?

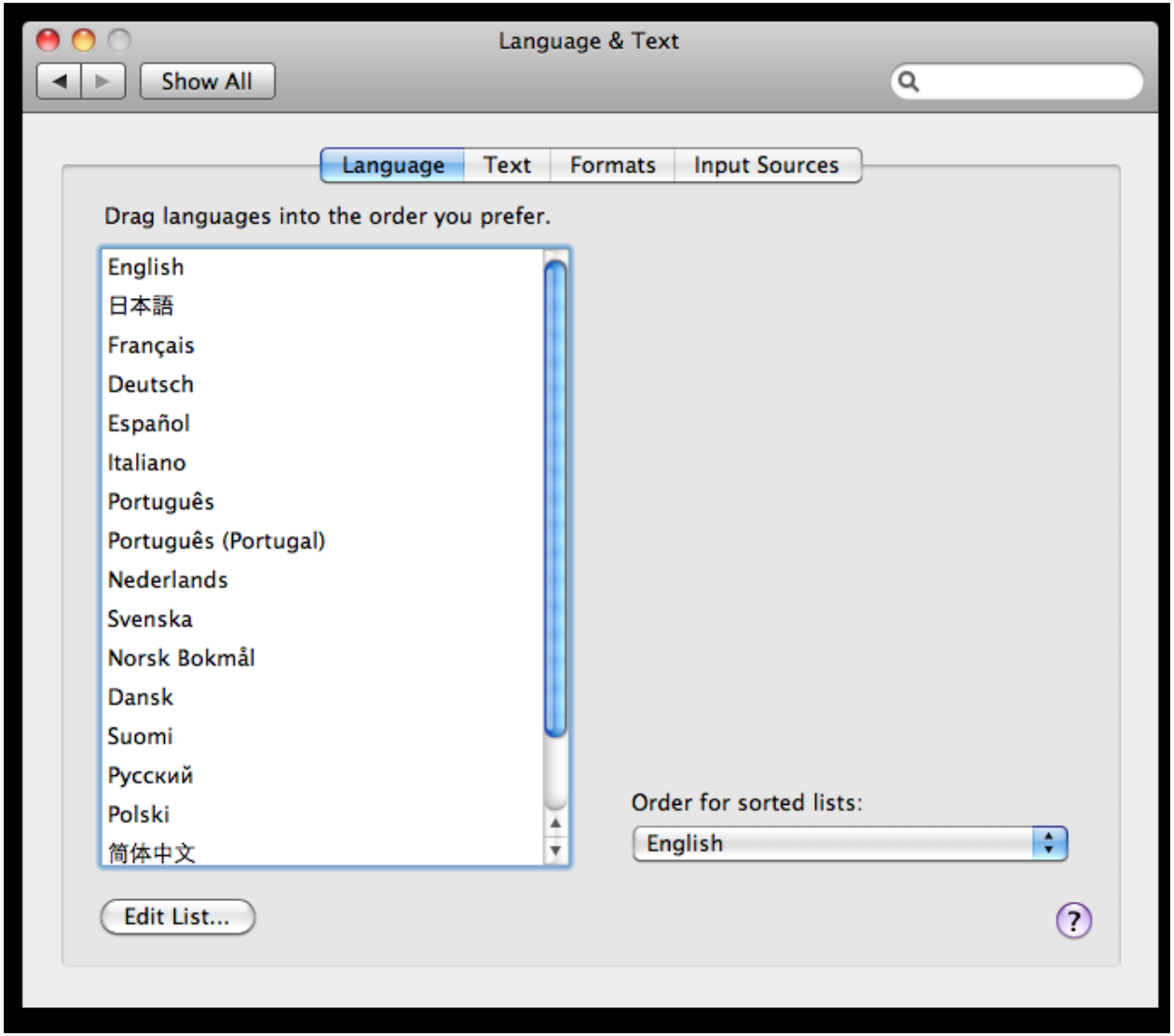
Defining language death

- Loss of function (other languages take over entire functional areas)
- Loss of prestige
- Loss of competence (emergence of semi-speakers) Menomini Bloomfield 1927, Gaelic Dorian 1981, Dyrbal Schmidt 1985

In the digital age:

- Loss of function performed digitally that is, increasingly, every function, from day to day communication (texting, email, ...) to commerce, official business, ...
- Loss of prestige If it's not on the web, it doesn't exist
- Loss of competence Can you raise a digital native in your language?

In the comfort zone



Full locale support, fonts, spellchecker, dictionaries, NLP tools

Must be FOSS – if it cannot be torrented it doesn't exist

Vital languages

- No wikipedia, no survival People know this – currently 133 proposals in incubator stage
- *But how good is a WP?*
- Estimate character entropy of language e.g. based on length of parallel texts. Filter out pages with longest paragraph shorter than 450 German chars *because these are weak/fake pages*
- Proportion of remaining pages gives real ratio, total (normalized) character count of real pages is adjusted WP size
- Potemkin wikipadias (e.g. Volapük) contribute nothing to survival
- Currently less than a hundred V+C

Vital languages

- No wikipedia, no survival People know this – currently 133 proposals in incubator stage
- *But how good is a WP?*
- Estimate character entropy of language e.g. based on length of parallel texts. Filter out pages with longest paragraph shorter than 450 German chars *because these are weak/fake pages*
- Proportion of remaining pages gives real ratio, total (normalized) character count of real pages is adjusted WP size
- Potemkin wikipadias (e.g. Volapük) contribute nothing to survival
- Currently less than a hundred V+C

Vital languages

- No wikipedia, no survival People know this – currently 133 proposals in incubator stage
- *But how good is a WP?*
- Estimate character entropy of language e.g. based on length of parallel texts. Filter out pages with longest paragraph shorter than 450 German chars *because these are weak/fake pages*
- Proportion of remaining pages gives real ratio, total (normalized) character count of real pages is adjusted WP size
- Potemkin wikipadias (e.g. Volapük) contribute nothing to survival
- Currently less than a hundred V+C

Vital languages

- No wikipedia, no survival People know this – currently 133 proposals in incubator stage
- *But how good is a WP?*
- Estimate character entropy of language e.g. based on length of parallel texts. Filter out pages with longest paragraph shorter than 450 German chars *because these are weak/fake pages*
- Proportion of remaining pages gives real ratio, total (normalized) character count of real pages is adjusted WP size
- Potemkin wikipadias (e.g. Volapük) contribute nothing to survival
- Currently less than a hundred V+C

Vital languages

- No wikipedia, no survival People know this – currently 133 proposals in incubator stage
- *But how good is a WP?*
- Estimate character entropy of language e.g. based on length of parallel texts. Filter out pages with longest paragraph shorter than 450 German chars *because these are weak/fake pages*
- Proportion of remaining pages gives real ratio, total (normalized) character count of real pages is adjusted WP size
- Potemkin wikipadias (e.g. Volapük) contribute nothing to survival
- Currently less than a hundred V+C

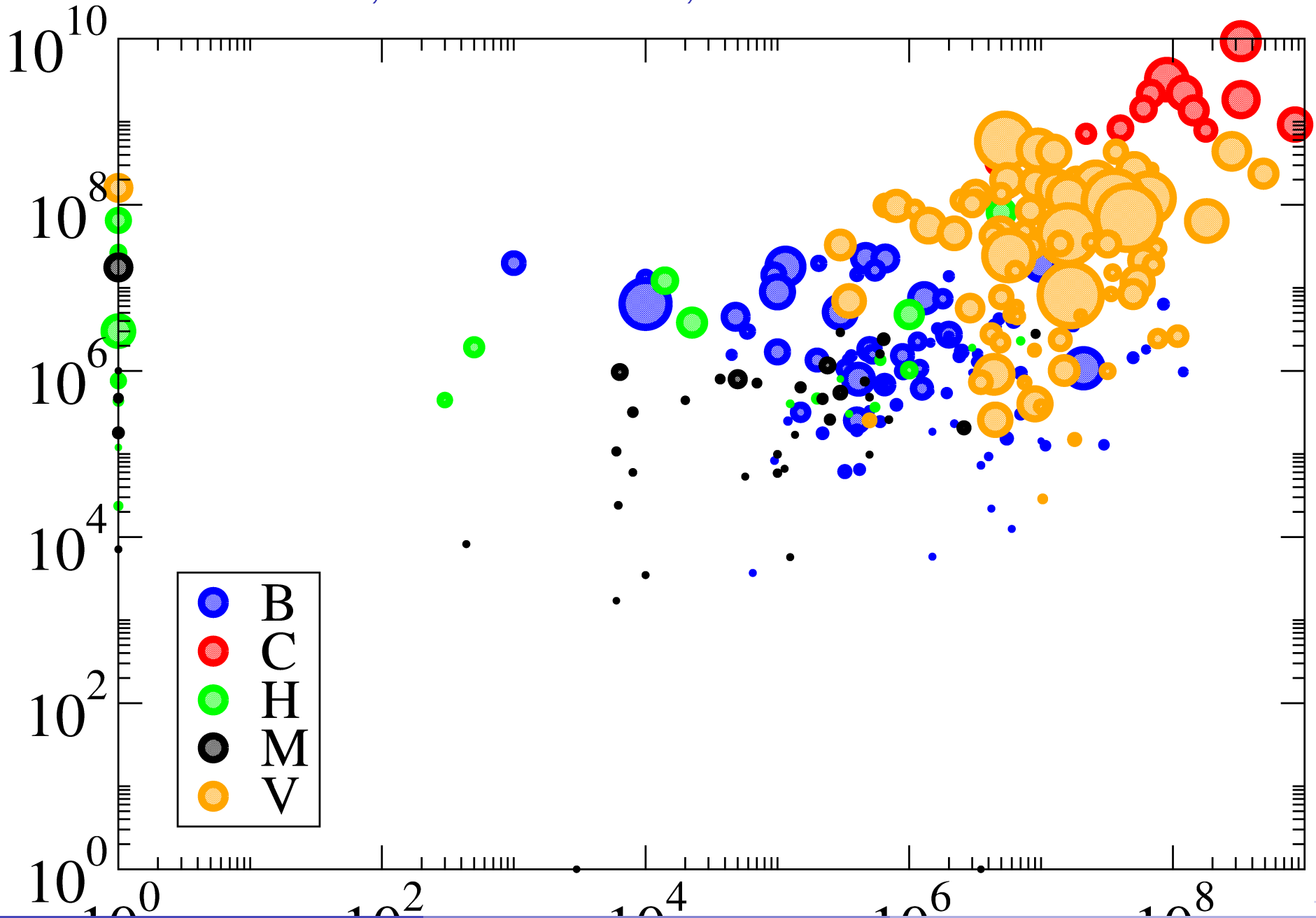
Vital languages

- No wikipedia, no survival People know this – currently 133 proposals in incubator stage
- *But how good is a WP?*
- Estimate character entropy of language e.g. based on length of parallel texts. Filter out pages with longest paragraph shorter than 450 German chars *because these are weak/fake pages*
- Proportion of remaining pages gives real ratio, total (normalized) character count of real pages is adjusted WP size
- Potemkin wikipadias (e.g. Volapük) contribute nothing to survival
- Currently less than a hundred V+C

Vital languages

- No wikipedia, no survival People know this – currently 133 proposals in incubator stage
- *But how good is a WP?*
- Estimate character entropy of language e.g. based on length of parallel texts. Filter out pages with longest paragraph shorter than 450 German chars *because these are weak/fake pages*
- Proportion of remaining pages gives real ratio, total (normalized) character count of real pages is adjusted WP size
- Potemkin wikipadias (e.g. Volapük) contribute nothing to survival
- Currently less than a hundred V+C

Community, WP size, real ratio



A four-way classification of languages C,V,H,M

C **In the comfort zone:** Wherever humanity goes, this language goes with them. 16 lgs, real ratio 0.36±0.10, #speakers 145m, WP 1.6g chars

V **Vital:** Significant digital community generating online material 83 lgs, real ratio 0.36±0.19, #speakers 31m, WP 96m chars

B **Borderline:** May yet make the transition to the digital age. 90 lgs, real ratio 0.15±0.13, #speakers 7m, WP 3.8m chars

- as a vital language B → V
- as a read-only carrier of cultural heritage B → H

H **Heritage:** 22 lgs, real ratio 0.14±0.13, #speakers 870k, WP 9.5m chars

M **Moribund or dead:** Digital natives cannot be raised. 41 lgs, real ratio 0.05±0.06, #speakers 840k, WP 970k chars

A four-way classification of languages C,V,H,M

C **In the comfort zone:** Wherever humanity goes, this language goes with them. 16 lgs, real ratio 0.36±0.10, #speakers 145m, WP 1.6g chars

V **Vital:** Significant digital community generating online material 83 lgs, real ratio 0.36±0.19, #speakers 31m, WP 96m chars

B **Borderline:** May yet make the transition to the digital age. 90 lgs, real ratio 0.15±0.13, #speakers 7m, WP 3.8m chars

- as a vital language B → V
- as a read-only carrier of cultural heritage B → H

H **Heritage:** 22 lgs, real ratio 0.14±0.13, #speakers 870k, WP 9.5m chars

M **Moribund or dead:** Digital natives cannot be raised. 41 lgs, real ratio 0.05±0.06, #speakers 840k, WP 970k chars

A four-way classification of languages C,V,H,M

- C **In the comfort zone:** Wherever humanity goes, this language goes with them. 16 lgs, real ratio 0.36±0.10, #speakers 145m, WP 1.6g chars
- V **Vital:** Significant digital community generating online material 83 lgs, real ratio 0.36±0.19, #speakers 31m, WP 96m chars
- B **Borderline:** May yet make the transition to the digital age. 90 lgs, real ratio 0.15±0.13, #speakers 7m, WP 3.8m chars
  - as a vital language B → V
  - as a read-only carrier of cultural heritage B → H
- H **Heritage:** 22 lgs, real ratio 0.14±0.13, #speakers 870k, WP 9.5m chars
- M **Moribund or dead:** Digital natives cannot be raised. 41 lgs, real ratio 0.05±0.06, #speakers 840k, WP 970k chars

A four-way classification of languages C,V,H,M

C **In the comfort zone:** Wherever humanity goes, this language goes with them. 16 lgs, real ratio 0.36±0.10, #speakers 145m, WP 1.6g chars

V **Vital:** Significant digital community generating online material 83 lgs, real ratio 0.36±0.19, #speakers 31m, WP 96m chars

B **Borderline:** May yet make the transition to the digital age. 90 lgs, real ratio 0.15±0.13, #speakers 7m, WP 3.8m chars

- as a vital language B → V
- as a read-only carrier of cultural heritage B → H

H **Heritage:** 22 lgs, real ratio 0.14±0.13, #speakers 870k, WP 9.5m chars

M **Moribund or dead:** Digital natives cannot be raised. 41 lgs, real ratio 0.05±0.06, #speakers 840k, WP 970k chars

A four-way classification of languages C,V,H,M

C **In the comfort zone:** Wherever humanity goes, this language goes with them. 16 lgs, real ratio 0.36±0.10, #speakers 145m, WP 1.6g chars

V **Vital:** Significant digital community generating online material 83 lgs, real ratio 0.36±0.19, #speakers 31m, WP 96m chars

B **Borderline:** May yet make the transition to the digital age. 90 lgs, real ratio 0.15±0.13, #speakers 7m, WP 3.8m chars

- as a vital language B → V
- as a read-only carrier of cultural heritage B → H

H **Heritage:** 22 lgs, real ratio 0.14±0.13, #speakers 870k, WP 9.5m chars

M **Moribund or dead:** Digital natives cannot be raised. 41 lgs, real ratio 0.05±0.06, #speakers 840k, WP 970k chars

A four-way classification of languages C,V,H,M

C **In the comfort zone:** Wherever humanity goes, this language goes with them. 16 lgs, real ratio 0.36±0.10, #speakers 145m, WP 1.6g chars

V **Vital:** Significant digital community generating online material 83 lgs, real ratio 0.36±0.19, #speakers 31m, WP 96m chars

B **Borderline:** May yet make the transition to the digital age. 90 lgs, real ratio 0.15±0.13, #speakers 7m, WP 3.8m chars

- as a vital language B → V
- as a read-only carrier of cultural heritage B → H

H **Heritage:** 22 lgs, real ratio 0.14±0.13, #speakers 870k, WP 9.5m chars

M **Moribund or dead:** Digital natives cannot be raised. 41 lgs, real ratio 0.05±0.06, #speakers 840k, WP 970k chars

A four-way classification of languages C,V,H,M

C **In the comfort zone:** Wherever humanity goes, this language goes with them. 16 lgs, real ratio 0.36±0.10, #speakers 145m, WP 1.6g chars

V **Vital:** Significant digital community generating online material 83 lgs, real ratio 0.36±0.19, #speakers 31m, WP 96m chars

B **Borderline:** May yet make the transition to the digital age. 90 lgs, real ratio 0.15±0.13, #speakers 7m, WP 3.8m chars

- as a vital language B → V
- as a read-only carrier of cultural heritage B → H

H **Heritage:** 22 lgs, real ratio 0.14±0.13, #speakers 870k, WP 9.5m chars

M **Moribund or dead:** Digital natives cannot be raised. 41 lgs, real ratio 0.05±0.06, #speakers 840k, WP 970k chars

A four-way classification of languages C,V,H,M

C **In the comfort zone:** Wherever humanity goes, this language goes with them. 16 lgs, real ratio 0.36±0.10, #speakers 145m, WP 1.6g chars

V **Vital:** Significant digital community generating online material 83 lgs, real ratio 0.36±0.19, #speakers 31m, WP 96m chars

B **Borderline:** May yet make the transition to the digital age. 90 lgs, real ratio 0.15±0.13, #speakers 7m, WP 3.8m chars

- as a vital language B → V
- as a read-only carrier of cultural heritage B → H

H **Heritage:** 22 lgs, real ratio 0.14±0.13, #speakers 870k, WP 9.5m chars

M **Moribund or dead:** Digital natives cannot be raised. 41 lgs, real ratio 0.05±0.06, #speakers 840k, WP 970k chars

Borderline languages

- No community, no survival The WP language policy states that at least five active users must edit that language regularly before a test project will be considered successful
- A group of enthusiasts can do wonders, but cannot sustain a lively community. As the digital death of Gaelic, Nynorsk, etc makes clear, the communities are voting with their smartphones
- Passive (read only) web presence (lexicons, classical literature, news services) is no substitute for active use in a broad variety of two-way contexts (social networks, business/commerce, live literature/blogs, etc) Heritage preservation has *huge* value!
- For any language pair, Google Translate likes to see gigaword monolingual corpora and megaword parallel text Reasonable goal for vitalization projects

Borderline languages

- No community, no survival The WP language policy states that at least five active users must edit that language regularly before a test project will be considered successful
- A group of enthusiasts can do wonders, but cannot sustain a lively community. As the digital death of Gaelic, Nynorsk, etc makes clear, the communities are voting with their smartphones
- Passive (read only) web presence (lexicons, classical literature, news services) is no substitute for active use in a broad variety of two-way contexts (social networks, business/commerce, live literature/blogs, etc) Heritage preservation has *huge* value!
- For any language pair, Google Translate likes to see gigaword monolingual corpora and megaword parallel text Reasonable goal for vitalization projects

Borderline languages

- No community, no survival The WP language policy states that at least five active users must edit that language regularly before a test project will be considered successful
- A group of enthusiasts can do wonders, but cannot sustain a lively community. As the digital death of Gaelic, Nynorsk, etc makes clear, the communities are voting with their smartphones
- Passive (read only) web presence (lexicons, classical literature, news services) is no substitute for active use in a broad variety of two-way contexts (social networks, business/commerce, live literature/blogs, etc) Heritage preservation has *huge* value!
- For any language pair, Google Translate likes to see gigaword monolingual corpora and megaword parallel text Reasonable goal for vitalization projects

Borderline languages

- No community, no survival The WP language policy states that at least five active users must edit that language regularly before a test project will be considered successful
- A group of enthusiasts can do wonders, but cannot sustain a lively community. As the digital death of Gaelic, Nynorsk, etc makes clear, the communities are voting with their smartphones
- Passive (read only) web presence (lexicons, classical literature, news services) is no substitute for active use in a broad variety of two-way contexts (social networks, business/commerce, live literature/blogs, etc) Heritage preservation has *huge* value!
- For any language pair, Google Translate likes to see gigaword monolingual corpora and megaword parallel text Reasonable goal for vitalization projects

Borderline languages

- No community, no survival The WP language policy states that at least five active users must edit that language regularly before a test project will be considered successful
- A group of enthusiasts can do wonders, but cannot sustain a lively community. As the digital death of Gaelic, Nynorsk, etc makes clear, the communities are voting with their smartphones
- Passive (read only) web presence (lexicons, classical literature, news services) is no substitute for active use in a broad variety of two-way contexts (social networks, business/commerce, live literature/blogs, etc) Heritage preservation has *huge* value!
- For any language pair, Google Translate likes to see gigaword monolingual corpora and megaword parallel text Reasonable goal for vitalization projects

Policy implications

- Different support for different stages. C languages can take care of themselves, V languages need comfort-enabler projects, B languages need vitalization *or* digital preservation projects
- EU has rich diversity experience and surprisingly deep expertise in languages outside its borders Use it or lose it
- Make sure the basic tools are built for all digitally viable languages – FOSS tools exist for many preservation projects (e.g. Nynorsk, Coptic) while lacking for vital languages (e.g Serbian)
- National projects need to make their corpora not just searchable but also downloadable by ROAMing (randomize, omit, anonymize, mix)

Policy implications

- Different support for different stages. C languages can take care of themselves, V languages need comfort-enabler projects, B languages need vitalization *or* digital preservation projects
- EU has rich diversity experience and surprisingly deep expertise in languages outside its borders Use it or lose it
- Make sure the basic tools are built for all digitally viable languages – FOSS tools exist for many preservation projects (e.g. Nynorsk, Coptic) while lacking for vital languages (e.g Serbian)
- National projects need to make their corpora not just searchable but also downloadable by ROAMing (randomize, omit, anonymize, mix)

Policy implications

- Different support for different stages. C languages can take care of themselves, V languages need comfort-enabler projects, B languages need vitalization *or* digital preservation projects
- EU has rich diversity experience and surprisingly deep expertise in languages outside its borders Use it or lose it
- Make sure the basic tools are built for all digitally viable languages – FOSS tools exist for many preservation projects (e.g. Nynorsk, Coptic) while lacking for vital languages (e.g Serbian)
- National projects need to make their corpora not just searchable but also downloadable by ROAMing (randomize, omit, anonymize, mix)

Policy implications

- Different support for different stages. C languages can take care of themselves, V languages need comfort-enabler projects, B languages need vitalization *or* digital preservation projects
- EU has rich diversity experience and surprisingly deep expertise in languages outside its borders Use it or lose it
- Make sure the basic tools are built for all digitally viable languages – FOSS tools exist for many preservation projects (e.g. Nynorsk, Coptic) while lacking for vital languages (e.g Serbian)
- National projects need to make their corpora not just searchable but also downloadable by ROAMing (randomize, omit, anonymize, mix)

Policy implications

- Different support for different stages. C languages can take care of themselves, V languages need comfort-enabler projects, B languages need vitalization *or* digital preservation projects
- EU has rich diversity experience and surprisingly deep expertise in languages outside its borders Use it or lose it
- Make sure the basic tools are built for all digitally viable languages – FOSS tools exist for many preservation projects (e.g. Nynorsk, Coptic) while lacking for vital languages (e.g Serbian)
- National projects need to make their corpora not just searchable but also downloadable by ROAMing (randomize, omit, anonymize, mix)