Digital Language Extinction as a Challenge for the Multilingual Web

Georg Rehm

Network Manager META-NET
DFKI, Berlin, Germany

georg.rehm@dfki.de

Multilingual Web Workshop 2014: New Horizons for the Multilingual Web
Madrid, Spain – May 8, 2014

Co-funded by the 7th Framework Programme and the ICT Policy Support Programme of the European Commission through the contracts T4ME, CESAR, METANET4U, META-NORD (grant agreements no. 249119, 271022, 270893, 270899).

Digital Language Extinction

META-NET

- Many smaller languages are experiencing problems digitally:
  - Loss of function – other languages take over entire functional areas such as, e.g., texting, email, search, e-commerce etc.
  - Loss of prestige – if it's not on the web, the languages doesn't exist
  - Loss of competence – can you raise a digital native in your language?
- Andras Kornai's classification – corresponds to the amount of digital communication in that language:
  1. digitally thriving languages (comfort zone languages)
  2. vital languages
  3. heritage languages
  4. still/moribund/dead languages

  potentially facing digital extinction …
- Implications for the European/global multilingual web?

- Network of Excellence dedicated to fostering the technological foundations of the European multilingual information society.
- Projects: T4ME, CESAR, METANET4U, META-NORD.
- First funded phase ended on Jan. 31, 2013; new projects such as, e.g., QTLaunchPad and QTLeap are contributing.
- All EU member states and several non-member states covered.
- META-NET: **60** research centres in **34** European countries.

http://www.meta-net.eu/members

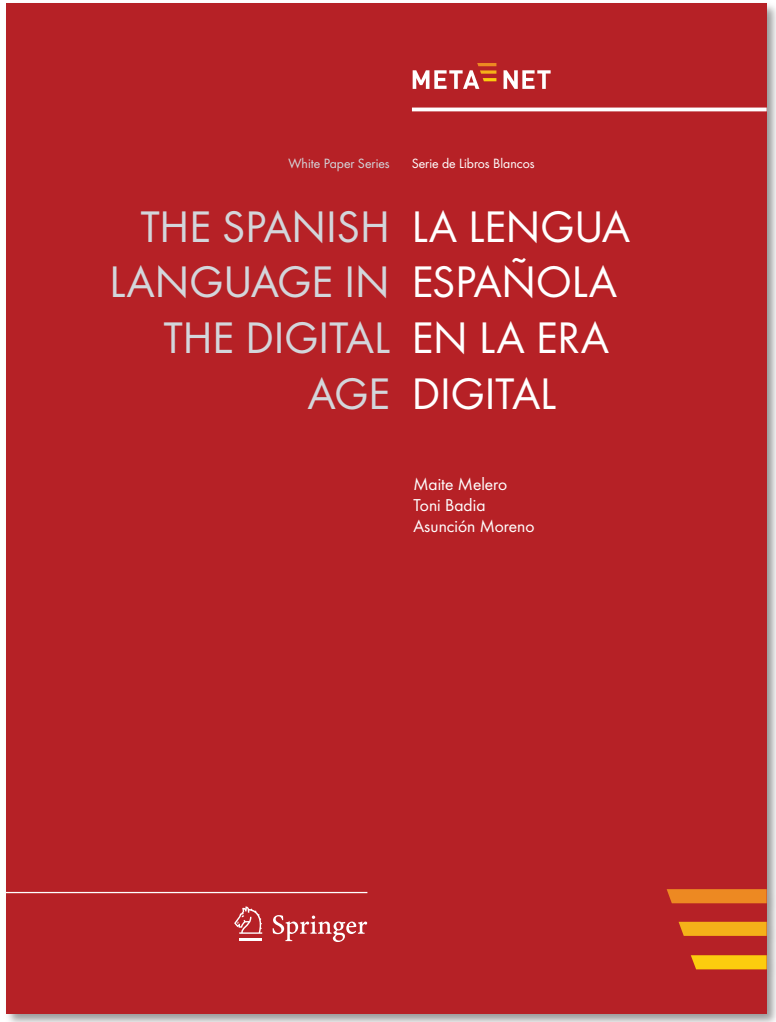
META-VISION: Building a community with a shared vision and strategic research agenda

META-SHARE: Building an open resource exchange infrastructure

META-RESEARCH: Building bridges to neighbouring technology fields

Language White Paper Series

- "Europe's Languages in the Digital Age"
- Series covers 31 languages in 31 volumes.
- Reports on the state of our languages in the digital age and the level of support through language technology.
- >2 years in the making.
- >215 experts as contributors.
- >8.000 copies distributed to politicians and journalists.



- Basque
- Bulgarian*
- Catalan
- Croatian*
- Czech*
- Danish*
- Dutch*
- English*
- Estonian*
- Finnish*
- French*
- Galician
- German*
- Greek*
- Hungarian*
- Icelandic
- Irish*
- Italian*
- Latvian*
- Lithuanian*
- Maltese*
- Norwegian
- Polish*
- Portuguese*
- Romanian*
- Serbian
- Slovak*
- Slovene*
- Spanish*
- Swedish*
- Welsh

* Official EU language

Cross-Lingual Comparison

- 1. **Machine Translation** 2. **Text Analytics**
  3. **Speech Processing/Synthesis** 4. **Language Resources**
- Ranking: from *excellent LT support* to *weak/no support*.
- Cross-lingual comparison discussed and finalised at a network meeting with representatives of all languages (Oct., 2011).

	excellent	good	moderate	fragmentary	weak or no support through LT
MT		English	French, Spanish	Catalan, Dutch, German, Hungarian, Italian, Polish, Romanian	Basque, Bulgarian, Croatian, Czech, Danish, Estonian, Finnish, Galician, Greek, Icelandic, Irish, Latvian, Lithuanian, Maltese, Norwegian, Portuguese, Serbian, Slovak, Slovene, Swedish, Welsh

	excellent	good	moderate	fragmentary	weak or no support through LT
Text Analytics		English	Dutch, French, German, Italian, Spanish	Basque, Bulgarian, Catalan, Czech, Danish, Finnish, Galician, Greek, Hungarian, Norwegian, Polish, Portuguese, Romanian, Slovak, Slovene, Swedish	Croatian, Estonian, Icelandic, Irish, Latvian, Lithuanian, Maltese, Serbian, Welsh

	excellent	good	moderate	fragmentary	weak or no support through LT
Speech		English	Czech, Dutch, Finnish, French, German, Italian, Portuguese, Spanish	Basque, Bulgarian, Catalan, Danish, Estonian, Galician, Greek, Hungarian, Irish, Norwegian, Polish, Serbian, Slovak, Slovene, Swedish	Croatian, Icelandic, Latvian, Lithuanian, Maltese, Romanian, Welsh

	excellent	good	moderate	fragmentary	weak or no support through LT
Resources		English	Czech, Dutch, French, German, Hungarian, Italian, Polish, Spanish, Swedish	Basque, Bulgarian, Catalan, Croatian, Danish, Estonian, Finnish, Galician, Greek, Norwegian, Portuguese, Romanian, Serbian, Slovak, Slovene	Icelandic, Irish, Latvian, Lithuanian, Maltese, Welsh

Observations and Results

- When it comes to technology support, there are *massive differences* between Europe's languages and technology areas.
- Support for English is ahead of any other language.
- But: even support for English is *far* from being perfect.
- Several languages get the weakest score in *all four areas* (e.g., Icelandic, Latvian, Lithuanian, Maltese)!

Digital Language Extinction!

- **"At Least 21 European Languages in Danger of Digital Extinction!"**
- Press release on European Day of Languages (Sept. 26, 2012).
- Huge global interest in the topic and our key findings!
- 600+ mentions in the press.
- News from 40+ countries in 35+ different languages.
- 20+ television reports and 30+ broadcast interviews (radio, tv) with META-NET representatives.
- Two Parliamentary Questions in the EP on the "digital extinction of languages" topic.

Ord. Forskere arbejder på at forbedre danske oversættelser på internettet.

Dårlig sprogteknologi truer dansk på nettet

Af Jens Ejsing
// ejs@berlingske.dk

Det danske sprog har det svært i den digitale verden.

Det konstaterer danske sprogforskere- og eksperter i forbindelse med den nye internationale undersøgelse META-NET, der ser nærmere på, hvordan en lang række mindre, europæiske sprog som dansk klarer sig i den digitale verden.

Forskerne fra bl.a. Københavns Universitet og Dansk Sprogtænkt når frem til, at dansk i fremtiden kan få det endnu sværere i den digitale verden, fordi Google Translate, GPSer, applikationer til smartphones og andre sprogteknologiske programmer ikke i tilstrækkelig grad formår at behandle de mange nuancer i det danske sprog.

Professor i sprogteknologi på Københavns Universitet, Bolette Sandford Pedersen, mener, at der er brug for en slags digital dansk sprogbank fyldt med data, så bl.a. oversættelser bliver så præcise og gode som muligt. Med hjælp fra sprogbanken kan forskere ifølge professoren hjælpe virksomheder med at forbedre programmer, der skal håndtere sproglig viden om bl.a. maskinoversættelse, talegenkendelse og informationssøgning.

Dermed vil der blive længere mellem fejlagtige oversættelser, som når »hånd olie på panden« med Google Translate bliver til »pour oil on the forehead« på engelsk. Oversættelser, der er i værste fald er så upræcise, at danskere ender med at fravælge deres eget sprog i den digitale verden.

Sprughjælp til virksomheder

Hun anerkender dog, at »teknologien til automatiske oversættelser på mange måder er fantastisk«.

»Den er bare ikke god nok, når det gælder dansk,« siger hun.

»Det er som om, at vi i et vist omfang lægger det i hænderne på Google eller andre virksomheder at afgøre, om dansk skal behandles godt nok eller ej. Men det danske marked er ikke stort for dem. Spørgsmålet er derfor, om vi ikke i højere grad selv skal gøre noget for at sikre, at det fornødne datamateriale er til rådighed, så vi får gode oversættelser og anden god sprogteknologi. Det kunne f.eks. være ved, at vi giorde en indsats for at få oprettet en sprogbank med en masse beriget materiale om dansk.«

»Hvis vi hele tiden oplever, at oversættelser er behæftede med fejl, tør vi ikke stole på dem,« siger hun og understreger, at »fejlagtige oversættelser kan føre til store misforståelser«.

Ifølge Dansk Sprognævns direktør, Sabine Kirchmeier-Andersen, kan dårlig sprogteknologi have konsekvenser for mange danskere, der ikke er så gode til engelsk.

»Hvis vi har ambitioner om at bruge det danske sprog i fremtidens teknologiske univers, skal der gøres en indsats nu for at fastholde eksperttise og udbygge den viden, vi har,« mener hun:

»Ellers risikerer vi, at kun folk, der taler flydende engelsk, vil få gråede af de nye generationer af web-, tele- og robotteknologi, der er på vej.«

fakta

Sprog i Europa

■ Der er omkring 80 sprog i EU. For 21 af dem – også dansk – gælder det, at der er store sprogteknologiske mangler, når det gælder bl.a. maskinoversættelse, talegenkendelse og informationssøgning.

■ Ifølge en EU-undersøgelse køber et stigende antal europæiske internetbrugere varer eller tjenester på nettet, hvor det sprog, der bliver anvendt, ikke er deres eget. Det gælder over halvdelen af brugerne.

■ Over hver tredje anvender et fremmedsprog til at skrive mail eller indlæg på nettet.

THE SUNDAY TIMES
Date 30 September 2012
Page 16
eclips newspaper licensing agency

BRIEFING DIGITAL IRISH

Lost for words

The Irish language is one of dozens of European tongues in danger of electronic extinction due to lack of digitised content. How can Gaeilge jump the digital divide, asks **John Burns**

BACKGROUND
Some European languages lagging

Research has been carried out into how 30 of Europe's estimated 80 languages are coping in the digital age, and the conclusion is most are unlikely to survive it. Language technology (LT) researchers estimate that digital support for 21 of the tongues is "non-existent" or "weak". The study by Meta-Net, a European network of 60 research centres in 34 countries, concluded that Irish, Latvian, Maltese and Slovenian are "weakly" under-resourced, while some are completely neglected. Irish is among those at risk of being left behind, but "with a strong pedigree in translation, localisation and LT research, Ireland is in a strong position to develop such technology", according to the authors, including Dr John Judge of Dublin City University.

INTERACTION
Irish-language users do have digital tools

LT covers everything from translating web pages to operating a mobile phone with spoken commands. There is software support for the Irish language, for example Google, Microsoft and Apple products. Microsoft's Windows 7 and Windows MS Office has a spell-checker, although it hasn't been updated recently, and there are online tools for spelling and grammar such as Gaeilgeoir. "There are localised versions of word processors, internet browsers and other desktop tools as well as operating systems," said Judge. "These technologies are being used in the Irish-speaking community." However there is no automatic speech recognition in Irish, and "there is a long way to go" before such systems can be put in place, says the study.

TONGUE TWISTER
Irish has unique quirks

The Irish language has distinctive features which make speech- and text-based technologies tricky. Unlike most European languages, it has a verb-subject-object structure, and adjectives follow nouns. Irish has no words for "yes" or "no". To give a yes or no response, you must repeat the verb. There are two forms of "to be", instead of a standard spoken form, there are three distinct dialects — Donegal, Connemara and Munster. As the standard spoken form has no native speakers — being a compromise of the three dialects — it "does not suit the building of a speech-synthesis system, which has to be based on a specific speaker". There is also an absence of resources which text-based technologies need, such as pronunciation dictionaries.

SPEAK UP
Urgent work needed to help Irish

Irish is largely out of the funding loop because it does not represent a market that investors are interested in, the report points out. Microsoft, Google and IBM all have centres for linguistics, and are based in Ireland, "but do very little work on Irish language data", being an official EU language affords Irish some protection, according to Judge. "But we desperately need to start producing more core component tools like parsers, chunkers, speech recognisers and synthesisers," he said. "In order to do this we need the 'raw materials', ie linguistic data in a machine-usable form. There is a lack of continuity in the maintenance of what little we have, which means existing resources run the risk of becoming obsolete."

38 Πέμπτη 27 Σεπτεμβρίου 2012 ΕΛΕΥΘΕΡΟΣ ΤΥΠΟΣ

GREEKLISH
Η γλώσσα της αποξένωσης...

ΠΟΛΛΕΣ ΕΥΡΟΠΑΪΚΕΣ ΓΛΩΣΣΕΣ ΘΕΩΡΟΥΝΤΑΙ ΤΕΧΝΟΛΟΓΙΚΑ... ΞΕΠΕΡΑΣΜΕΝΕΣ

Με ψηφιακή εξαφάνιση κινδυνεύουν τα ελληνικά

ΕΛΕΝΗ ΒΕΡΤΟΥ
evertou@e-typos.com

Σ'την ψηφιακή εποχή δύο... χιλιάδες ελληνικά, όπως και άφοπτές άλλες ευρωπαϊκές γλώσσες, σύμφωνα με πανευρωπαϊκή έκθεση με την υπογραφή 200 και πλέον ειδικούς. Η συγκομιδή μελών δημοσιεύτηκε από το ινστιτούτο Meta-Net με αφορμή τη σύνοδο Ευρωπαϊκή Ημέρα Γλωσσών.

Για τις ανάγκες της έρευνας τους, γλωσσολόγοι από 54 κόμες της Γης, Ηνωμένο βασιλειάδηγήσεις τις διαθέσιμες γλωσσικές υπηρεσίες και δημιούργησαν ένα «Λευκό Βιβλίο» για κάθε ευρωπαϊκή γλώσσα. Στη μελέτη τους, οι ειδικοί ανέβασαν στο μπαλί άλλων τάσεων βασικά πλεονεκτικά εργαλεία, δηλαδή τον αναμφίβολο αυτόματος μετάφραστη, τη δυνατότητα φωνητικής αλληλεπίδρασης και φτιαχτές ανάλυσης κειμένου, ενώ ταυτόχρονα διαφαινόθηκε και η διαθεσιμότητα γλωσσικών πόρων ή πηγών.

Σε πρώτο φάσο εξέτασαν τις υπολειπόμενες ή ανεπαρκείς ή ανύπαρκτες κάποιες τις κάνουν μπαταρίες: ανίκανες, για παράδειγμα, η υπηρεσία του κυβέρνησε πληροφορικής Google Translate. Την ίδια ώρα, εξετάστηκαν τι αποκαταστάσεις των ελληνικών κρατούν με τις, ανακαινές τους, όπως για παράδειγμα η δυνατότητα τεχνολογίας ανάμεσα στις διάφορες ευρωπαϊκές γλώσσες. Το κάτω μέρος βρίσκεται η ελληνική γλώσσα, ενώ τα τελευταία δεκαετία και η ελληνική γλώσσα, ενώ σε λίγο καλύτερη μόρφο βρίσκονται οι ελληνικοί, η βουλγαρική, η ουγγρική και η πολωνική, που όπως αναφέρει η έρευνα έκανε «αποσπασματικές τεχνολογικές».

Αθέμιτα, χαρακτηρίζεται η υπαρκτή κρίση σε ολόκληρη, χωρίς λοιπή, ψηφιακή, σαλκή και ιστικνιά γλώσσα. Οι ενεργειές της υπεύθυνες ομάδες, Χαις Ολβόφρεν και Ντάβιντ Ρις, ανιχνεύουν τρεις κρίσιμες πτυχές στο υπαστήρι της γλωσσικής τεχνολογίας: Ανά «ιστανολογική αποκλεισμός» ανάμεσα στις γλώσσες, ένα ταυτόχρονο διαχωρισμό ανάμεσα στις γλώσσες, ένα τεχνολογικό χάσμα ανάμεσα στις γλώσσες.

Ο κίνδυνος μεταβλήτου κατακτά, όπως λέει αλλιώτα και λογική, η ουγγρική γλώσσα. Οι ουγγλόφωνοι κρίνουν ότι η κλιμάτο δυνατή κοινωνική υποστήριξη, κάτι το οποίο κινεί την παρατήρα εξέτασων στις γλώσσες. Ανά «ιστανολογική αποκλεισμός» ανδυνούνε περισσότερο ή πολυλαλή, η λιτανική, η λιθουανική και η ρωμανίζεστη γλώσσα, ενώ σε λίγο καλύτερη μόρφο βρίσκονται οι ελληνικοί, η βουλγαρική, η ουγγρική και η πολωνική, που όπως αναφέρει η έρευνα έκανε «αποσπασματικές τεχνολογικές».

Ο ανγγλόφωνοι χρήστες έχουν την καλύτερη δυνατή τεχνολογική υποστήριξη, γεγονός που ευνοεί την περαιτέρω εξάπλωση της γλώσσας

Ενεργοποιείς του λόγου Ερευνητικό Κέντρο Αθηνά, λέει στον «Ε.Τ.» «Η γλώσσα μου λέει ότι δεν θα ζήσει η ελληνική γλώσσα ή ότι κινδυνεύει με εξαφάνιση». Η ιδέα; εξηγεί ότι ένα υπάρχουν ανθρωποι που μιλούν, γράφουν και επικοινωνούν με μια γλώσσα, τότε αυτή θα συνεχίσει να υπάρχει. Είναι σημαντικό, όμως, να ξέρει ότι οι κρίσεις τη δυνατότητα αυτόματου σε νέοι ανθρωποι, όπως τα GPS τους, στα ελληνικά και αν ξανάβαλε τους γλωσσικά εργαλεία αυτές δύσκολα θα.. ανθίσουν στον ψηφιακό κόσμο τον 21ο αιώνα. Η Μαρία Γαβριηλίδου, μέλος της επιτροπικής ομάδας από τα λιτσιάτα κλαμπάν» ανδυνούνε περισσότερο ή πολυλαλή, η λιτανική, η λιθουανική και η ρωμανίζεστη γλώσσα.

Μέλεστο, κρίσιμο ζήτημα σε ωμφαιά εξαιρέτων μας γλώσσες είναι η ωμφαιά εξέτασων μας γλώσσες ή ωμφαιά εξέτασων μας γλώσσα χρήστων. Οι ανάστασε κιβόρνιας, οι Ευρωπαϊκή Ένωση και ο ιδιωτικός τομέας πρέπει να κρηφαιοδοτήσουν την ανάπτυξη αυτής της τεχνολογίας για όλες τις γλώσσες, ανεφέρει και συνεχίζει «Σε πρόταση; ομάδες πρέπει να φτιάξουν ανθρώπινα και ανεξάρτητα μεταξύ τους αυτά και να μην ικανοποιούνται με τα αγγλικά. ■

ΔΗΜΟΚΡΑΤΙΑ 26/09/2012 1:41 PM Page 49

η σημερινή | ΚΥΡΙΑΚΗ 30 ΣΕΠΤΕΜΒΡΙΟΥ 2012

Οι περισσότερες ευρωπαϊκές γλώσσες κινδυνεύουν με ψηφιακή εξαφάνιση

Τη γλώσσα μού... έχασαν

Πρέπει να εξασφαλιστεί ο εφοδιασμός των μικρότερων και λιγότερο πλούσιων -σε ψηφιακούς πόρους- γλωσσών με τις απαραίτητες βασικές τεχνολογίες

Η λίστα Γιστιμμάχου έχει καθιερωθεί ανάμεσα στα ενδιαφερόμενα μέρη της Ευρώπης σε η Ευρωπαϊκή Ένωση, σε ο Ευρωπαϊκό Κοινοβούλιο, αλλά σύμφωνα με μια νέα συγκεντρωτική-ανιχνευτική έκθεση, σε ο από τις 21 γλώσσες της Ευρώπης -μεταξύ των οποίων η ελληνική, η ιρλανδική, η λιθουανική, η ουγγρική, η πολωνική- δεν υπάρχει ικανοποιητικός εφοδιασμός ψηφιακών πόρων.

Η έρευνα φέρει τον κόδωνα κοίνωνα, ανάμεσα στις γλώσσες η ιρλανδική βρίσκεται σε τις περισσότερες ευρωπαϊκές γλώσσες είναι ελλιπής ή υπολείπεται ανάπτυκτη για τους χρήστες.

Τις έφερναν οι κοσκίνες

Σε ανάπτυξη, με τη χορηγό μιας ακόμη λιλιάν βίβλια (με τίτλο «Λευκός Γίγαντας» Ευρωπαϊκή Κοινωνία της Πληροφορίας), από τα ενιαίοτερα-ιστικά δίκτυα META-NET, το οποίο συνενώνει 60 ερευνητικά κέντρα σε 34 χώρες, ενσωμάτωνε ότι η γλώσσες που μελετούνε από κοινό μαρμί ομάδων κοίνωνίας, ανάμεσα στις οποίες η ιρλανδική, η λιθουανική, η Ουγγρική και η Πολωνική, ανάπτυξη κινδυνεύουν στον σύγχρονο ψηφιακό κόσμο.

Η έρευνα του META-NET, στην οποία συνέβαλαν κάποιοι ειδικοί και ιδιωτικά, αξιολογεί τον κίνδυνο για κάθε γλώσσα με βάση τέσσερα βασικά κριτήρια: τα ψηφιακά γλωσσικά επίπεδο; τις ύπαρκτη κρυφής γλώσσας, τη δυνατότητα φωνητικής αλληλεπίδρασης, τη δυνατότητα ψηφιακής ανάλυσης κειμένου και τη διαθεσιμότητα των σχετικών ψηφιακών γλωσσικών πόρων/πηγών.

Οι δυνατές

Σε γλώσσα με την υπάλληλη βιβλιοθήκη ψηφιακών πόρων είναι ικανοποιητικά η Δανική, που αποκαλείται τη σκανδιναβική κλίμακα τεχνολογική υποστήριξη (σε και ήρ τον κυλάνη δυνατή), ενγνώνε που δεκαπλάσιο την περιοχήα εξίσωσή της.

Ακολουθούν με ικανοποιητική ή μέτρια τεχνολογική υποστήριξη (ανάμεσα η Ολλανδική, η Γαλλική, η Γερμανική, η Γαλικιά και η Ισπανική, η Βλάχικη, η Βασκική, η Καταλανική, η Πολωνική, η Ουγγρική κ.ά.) κάποιοιστοιχεία στις γλώσσες με κατασπαρμένη ψηφιακή υποστήριξη, η οποία ακμάζει σαφώνεται πλείστες φορές κινδύνου ηρμε «ξεφτάνει».

Δραματικές διαφορές

Κλείνουν με τους σταμάτηστ τις μειλίτης Χαις Ολβόφρεν και Γκίμφρν Ρις, ανιχνεύουν δραματικές διαφορές στην υποστήριξη της γλωσσικής τεχνολογίας ανάμεσα στις ευρωπαϊκές γλώσσες και τεχνολογικές περιοχές. Τα φόνια μεταξύ μικρών και μεγάλων γλωσσών ιλιάσια και διαφόρεια. Πρέπει να εξασφαλίσουμε τον εφοδιασμό των μικρότερων και λιγότερο πλούσιων -σε ψηφιακούς πόρους- γλωσσών με τις απαραίτητες βασικές τεχνολογίες, ειδικές οι γλώσσες αυτές είναι καταδικασμένες σε ψηφιακούς πόρους γλωσσών με τις απαραίτητες βασικές τεχνολογίες, ειδικές οι γλώσσες αυτές είναι καταδικασμένες σε ψηφιακούς πόρους γλωσσών με τις απαραίτητες βασικές τεχνολογίες.

Σε ορισμένες περιπτώσεις η ανάπτυξη αυτήν τη βασικών γλωσσικών τεχνολογιών θα μπορούσε η ΕΕ η Ευρωπαϊκή Κοινοβουλή Κοινωνία της Πληροφορίας, από τα ενιαίοτερα-ιστικά δίκτυα META-NET, το οποίο συνενώνει 60 ερευνητικά κέντρα σε 34 χώρες, ενσωμάτωνε ότι η γλώσσες που μελετούνε από κοινό μαρμί ομάδων κοίνωνίας, ανάμεσα στις οποίες η ιρλανδική, η λιθουανική, η Ουγγρική και η Πολωνική, ανάπτυξη κινδυνεύουν στον σύγχρονο ψηφιακό κόσμο.

Η έρευνα του META-NET, στην οποία συνέβαλαν κάποιοιστοιχεία από κοινό ειδικοί, αξιολογεί τον κίνδυνο για κάθε γλώσσα με βάση τέσσερα βασικά κριτήρια: τα ψηφιακά γλωσσικά επίπεδο; τις ύπαρκτη κρυφής γλώσσας, τη δυνατότητα φωνητικής αλληλεπίδρασης, τη δυνατότητα ψηφιακής ανάλυσης κειμένου και τη διαθεσιμότητα των σχετικών ψηφιακών γλωσσικών πόρων/πηγών.

Τέσσερα μεγάλα κίνδυνα

Κλείνουν με τη νίκη μάλιστα, η Δανική, η Λετονική, η Λιθουανική και η Μαλτέζικη ανταμπαίνουν τον μεγαλύτερο κίνδυνο εξαφάνισης σε μια ευρωπαϊκή τεχνολογική κοινωνία, που ολοένα περισσότερο πρωιαθεί τη χρήση συγκεκριμένων γλωσσών και ιδίως τις γλώσσες αυτές. Η ίδια η Ιρλανδική, η Μαλτέζικη, η Λετονική, η Λιθουανική, η Ουγγρική και η Πολωνική, ανάπτυξη κινδυνεύουν στον σύγχρονο ψηφιακό κόσμο.

Η έρευνα του META-NET, στην οποία συνέβαλαν κάποιοιστοιχεία από κοινό ειδικοί, αξιολογεί τον κίνδυνο για κάθε γλώσσα με βάση τέσσερα βασικά κριτήρια: τα ψηφιακά γλωσσικά επίπεδο; τις ύπαρκτη κρυφής γλώσσας, τη δυνατότητα φωνητικής αλληλεπίδρασης, τη δυνατότητα ψηφιακής ανάλυσης κειμένου και τη διαθεσιμότητα των σχετικών ψηφιακών γλωσσικών πόρων/πηγών.

Update of the Study (2014)

- Study comprised 31 volumes/languages.
- Many languages missing! Need for extension – at least of the comparison.
- We invited three language community bodies to participate in the update:

  European Federation of National Institutions for Language (EFNIL)

  Network to Promote Linguistic Diversity (NPLD)

  Experts Committee of the European Language Charter (Council of Europe)

An Update and Extension of the META-NET Study "Europe's Languages in the Digital Age"

Georg Rehm[1], Hans Uszkoreit[1], Ido Dagan[2], Vartkes Goetcherian[3], Mehmet Ugur Dogan[4], Coskun Mermer[4], Tamás Varadi[5], Sabine Kirchmeier-Andersen[6], Gerhard Stickel[7], Meirion Prys Jones[8], Stefan Oeter[9], Sigve Gramstad[10]

META-NET
DFKI GmbH
Berlin, Germany[1]

META-NET
Bar-Ilan University
Tel Aviv, Israel[2]

META-NET
Arax Ltd
Luxembourg[3]

META-NET
Tübitak Bilgem
Gebze, Turkey[4]

EFNIL, META-NET
Hungarian Academy of Sciences
Budapest, Hungary[5]

EFNIL, META-NET
Danish Language Council
Copenhagen, Denmark[6]

EFNIL
Institut für Deutsche Sprache
Mannheim, Germany[7]

NPLD
Network to Promote Ling. Diversity
Cardiff, Wales[8]

Council of Europe, Com. of Experts
University of Hamburg
Hamburg, Germany[9]

Council of Europe, Com. of Experts
Bergen, Norway[10]

Abstract

This paper extends and updates the cross-language comparison of LT support for 30 European languages as published in the META-NET Language White Paper Series. The updated comparison confirms the original results and paints an alarming picture: it demonstrates that there are even more dramatic differences in LT support between the European languages.

Keywords: LR National/International Projects, Infrastructural/Policy Issues, Multilinguality, Machine Translation

1. Introduction and Overview

The multilingual setup of our European society imposes societal challenges on political, economic and social integration and inclusion, especially in the creation of the single digital market and unified information space targeted by the Digital Agenda (EC, 2010). Language technology is the missing piece of the puzzle, it is the key enabler and solution to boosting growth and strengthening Europe's competitiveness.

Recognising Europe's exceptional demand and opportunities, 60 leading research centres in 34 European countries joined forces in META-NET, a Network of Excellence dedicated to the technological foundations of a multilingual European information society. META-NET was partially supported through four projects funded by the EC: T4ME, CESAR, METANET4U and META-NORD. META-NET is forging the Multilingual Europe Technology Alliance (META) with more than 760 organisations and experts representing multiple stakeholders and signed collaboration agreements with more than 40 other projects and initiatives. META-NET's goal is monolingual, crosslingual and multilingual technology support for all European languages (Rehm and Uszkoreit, 2013). We recommend focusing on three priority research themes connected to application scenarios that will provide European R&D with the ability to compete with other markets and achieve benefits for European society and citizens as well as opportunities for our economy and future growth.

This paper extends and updates one important result of the work carried out within the META-VISION pillar of the initiative, the cross-language comparison of LT support for 30 European languages as published in the META-NET Language White Paper Series (Rehm and Uszkoreit, 2012).

2. The Language White Paper Series

Answering the question on the current state of a whole R&D field is difficult and complex. For LT nobody had collected these indicators and provided comparable reports for a substantial number of European languages yet. To arrive at a first comprehensive answer, META-NET prepared the Language White Paper Series "Europe's Languages in the Digital Age" (Rehm and Uszkoreit, 2012) that describes the current state of LT support for 30 European languages (including all 24 official EU languages). This undertaking had been in preparation with more than 200 experts since mid 2010 and was published in the summer of 2012. The study included a comparison of the support all languages receive in four areas: MT, speech, text analytics, language resources. The differences in technology support between the various languages and areas are dramatic and alarming. In the four areas, English is ahead of the other languages but even support for English is far from being perfect. While there are good quality software and resources available for a few larger languages and application areas, others, usually smaller languages, have substantial gaps. Many languages lack basic technologies for text

	excellent	good	moderate	fragmentary	weak or no support
MT		English	French, Spanish	Catalan, Dutch, German, Hungarian, Italian, Polish, Romanian	**Albanian**, **Asturian**, Basque, **Bosnian**, **Breton**, Bulgarian, Croatian, Czech, Danish, Estonian, Finnish, **Frisian**, **Friulian**, Galician, Greek, **Hebrew**, Icelandic, Irish, Latvian, **Limburgish**, Lithuanian, **Luxembourgish**, **Macedonian**, Maltese, Norwegian, **Occitan**, Portuguese, **Romany**, **Scots**, Serbian, Slovak, Slovene, Swedish, **Turkish**, **Vlax Romani**, Welsh, **Yiddish**
	excellent	good	moderate	fragmentary	weak or no support
Text Analytics		English	Dutch, French, German, **Hebrew**, Italian, Spanish	Basque, Bulgarian, Catalan, Czech, Danish, Finnish, Galician, Greek, Hungarian, Norwegian, Polish, Portuguese, Romanian, Slovak, Slovene, Swedish	**Albanian**, **Asturian**, **Bosnian**, **Breton**, Croatian, Estonian, **Frisian**, **Friulian**, Icelandic, Irish, Latvian, **Limburgish**, Lithuanian, **Luxembourgish**, **Macedonian**, Maltese, **Occitan**, **Romany**, **Scots**, Serbian, **Turkish**, **Vlax Romani**, Welsh, **Yiddish**
	excellent	good	moderate	fragmentary	weak or no support
Speech		English	Czech, Dutch, Finnish, French, German, Italian, Portuguese, Spanish	Basque, Bulgarian, Catalan, Danish, Estonian, Galician, Greek, Hungarian, Irish, Norwegian, Polish, Serbian, Slovak, Slovene, Swedish, **Turkish**	**Albanian**, **Asturian**, **Bosnian**, **Breton**, Croatian, **Frisian**, **Friulian**, **Hebrew**, Icelandic, Latvian, **Limburgish**, Lithuanian, **Luxembourgish**, **Macedonian**, Maltese, **Occitan**, Romanian, **Romany**, **Scots**, **Vlax Romani**, Welsh, **Yiddish**
	excellent	good	moderate	fragmentary	weak/no support
Resources		English	Czech, Dutch, French, German, Hungarian, Italian, Polish, Spanish, Swedish	Basque, Bulgarian, Catalan, Croatian, Danish, Estonian, Finnish, Galician, Greek, **Hebrew**, Norwegian, Portuguese, Romanian, Serbian, Slovak, Slovene	**Albanian**, **Asturian**, **Bosnian**, **Breton**, **Frisian**, **Friulian**, Icelandic, Irish, Latvian, **Limburgish**, Lithuanian, **Luxembourgish**, **Macedonian**, Maltese, **Occitan**, **Romany**, **Scots**, **Turkish**, **Vlax Romani**, Welsh, **Yiddish**

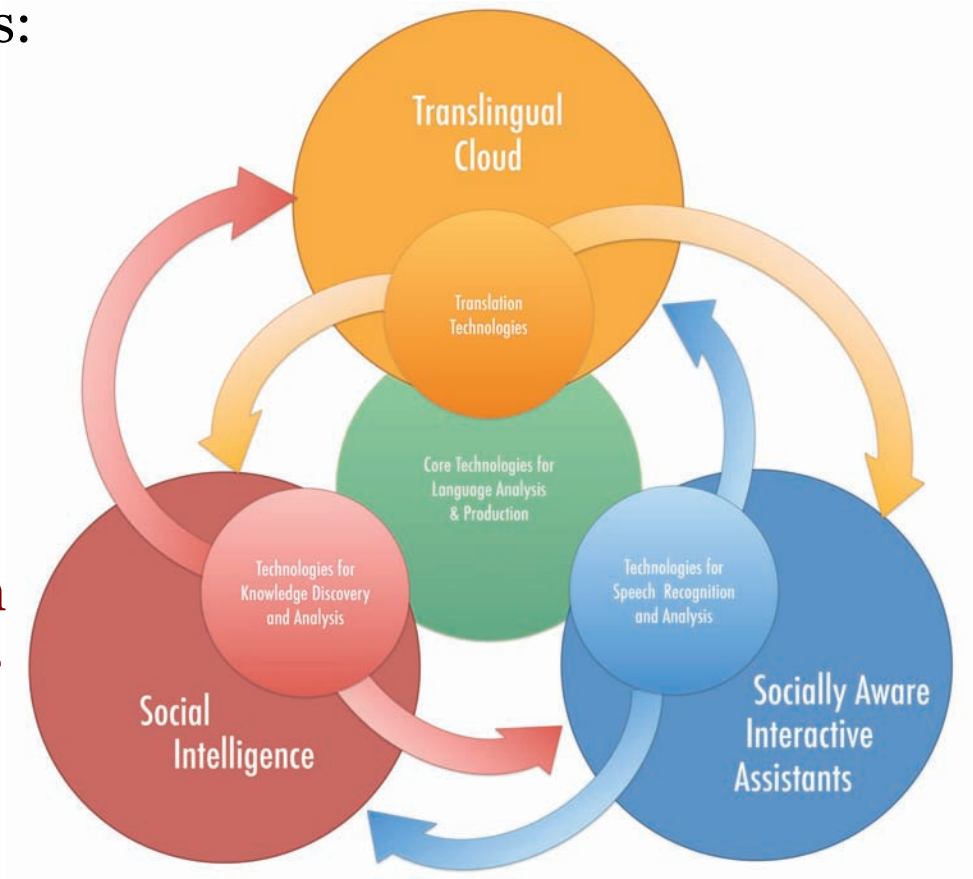
Strategic Research Agenda

- Addresses the problems we identified when preparing the white papers.
- Can put Europe ahead of its competitors in this technology area.
- 200 contributors; >2 years.

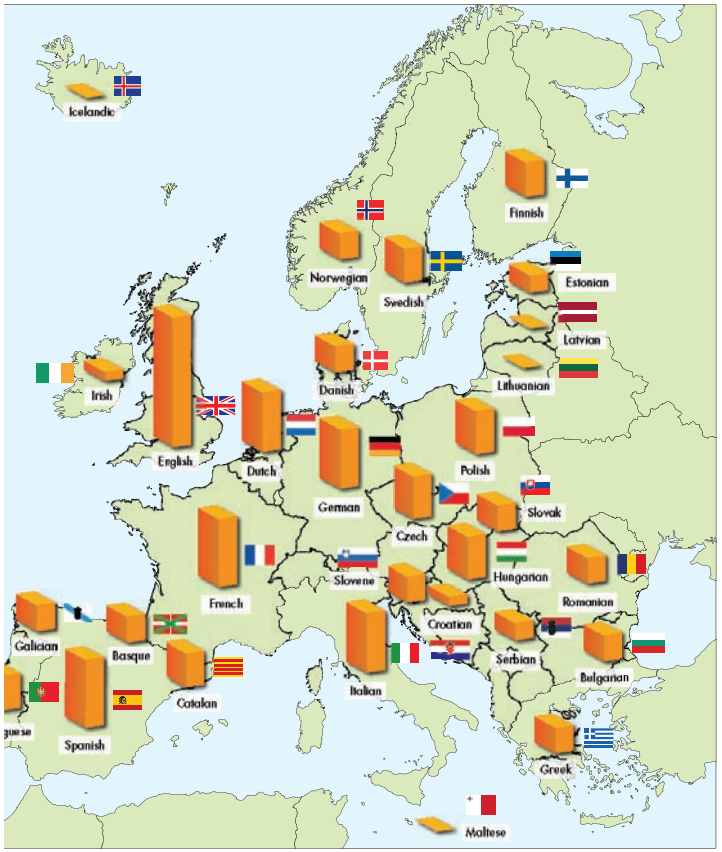
  54% industry; 46% research;
  4% (inter)national institutions.
- Presented and discussed at 90+ conferences and major workshops.
- Published & presented in early 2013.
- http://www.meta-net.eu/sra

Priority Research Themes

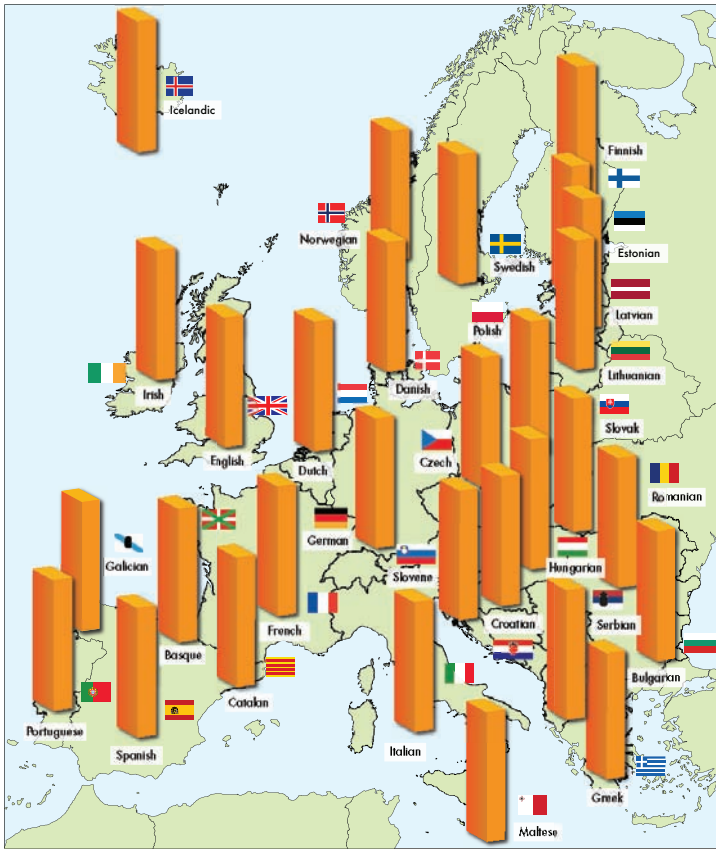
- Three priority research themes:
  - Translingual Cloud
  - Social Intelligence and e-Participation
  - Socially-Aware Interactive Assistants
- Two additional themes:
  - European Service Platform for Language Technologies
  - Core Technologies for Language Analysis and Production



Resources, Tools, Technologies



2013



2020

Summary and Conclusions

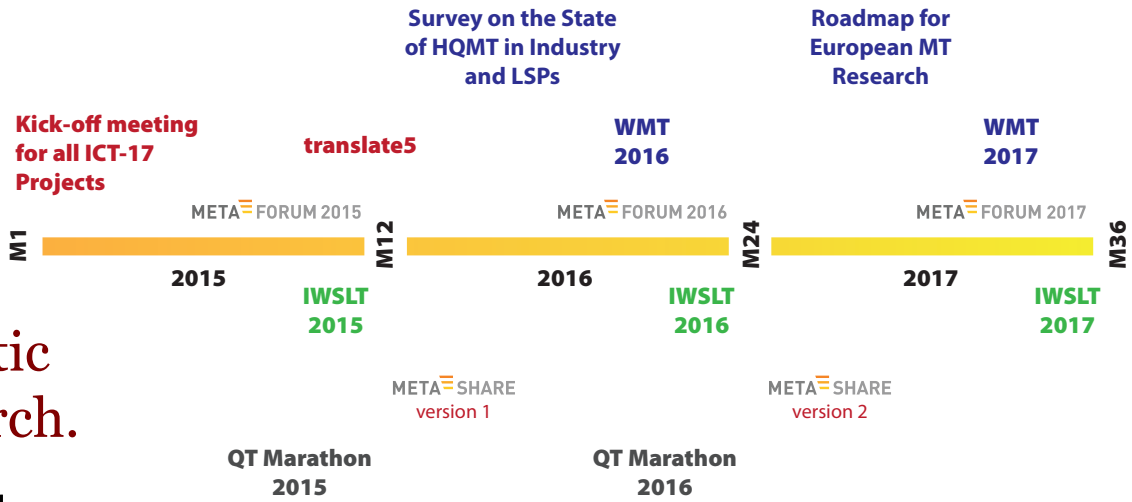
- Many languages suffer from insufficient technology support.
- These languages are threatened by digital extinction.
- Multilingual technologies are a key enabler to overcoming language barriers and a good means to preserving our languages.
- Policy implications! We need different support for Kornai's classes:
  - Only digitally thriving languages can take care of themselves!
  - Vital, heritage, still/moribund languages need our help!
  - Focus on fostering R&D for smaller, less-resourced languages as well as research and technology transfer between languages.
- We suggest setting up an interdisciplinary research effort so that Europe can overcome language barriers, benefit from language diversity and fight digital language extinction (see META-NET SRA).

Current Funding Opportunities

- Horizon 2020-ICT-17 ("Cracking the language barrier"): Focus on languages with *fragmentary* or *weak/no* support through LT. ☺
  - Official EU Member State languages only! 😐☹
  - Funding is limited: only 15M€ (out of 658M€ for ICT in 2015/2016). ☹☹
- CEF: MT is considered an obligatory component, core building block, key component of Europe's future digital infrastructure! ☺☺
  - Budget of only 4M€ in CEF-WP-2014 for procurement. ☹☹
  - Prep phase only! MT not considered ripe for production use just yet. 😐☹

ICT-17c Proposal: CRACKER

- CRACKER – Cracking the Language Barrier: Coordination, Evaluation and Resources for European MT Research
- CSA (budget: 1M€), 2015 to 2017.
- DFKI, CUNI, ELDA, FBK, ATH, UEDIN, USFD
- Pushes towards a systematic improvement of MT research.



- Nucleus of strategy towards HQ-MT is the future ICT-17 group of projects.
- Activities: evaluation campaigns (WMT, IWSLT), META-FORUM, META-SHARE, training (MT Marathons), coordination.

¡Muchas gracias!

- Digital language extinction is a serious threat for many languages.
- The Multilingual Web community can help by providing resources, tools, technologies for these languages.
- Together we can fight digital language extinction!

http://www.meta-net.eu
http://www.facebook.com/META.Alliance

Acknowledgements: This work would not have been possible without the dedication and commitment of our colleagues Aljoscha Burchardt, Kathrin Eichler, Tina Klüwer, Arle Lommel, Felix Sasaki and Hans Uszkoreit (all DFKI), the 60 member organisations of the META-NET network of excellence, the ca. 70 members of the Vision Groups, the ca. 30 members of the META Technology Council, the more than 200 authors of and contributors to the META-NET Language White Paper Series and the ca. 200 representatives from industry and research who contributed to the META-NET Strategic Research Agenda.