European Language Grid – META-FORUM 2020
Session 4: News from the Language Communities

- **Marko Turpeinen (1001 Lakes, Finland)**
- Walter Daelemans (University of Antwerp, NCC Belgium)
- Svetla Koeva (Bulgarian Academy of Sciences, NCC Bulgaria)
- Maciej Ogrodniczuk (Polish Academy of Sciences, NCC Poland)
- Marta Villegas (Barcelona Supercomputing Center, NCC Spain)
- François Yvon (LIMSI/CNRS, NCC France)

“LANGUAGE PLATFORM” NATIONAL INITIATIVE FROM FINLAND

META-FORUM 2020

02.12.2020

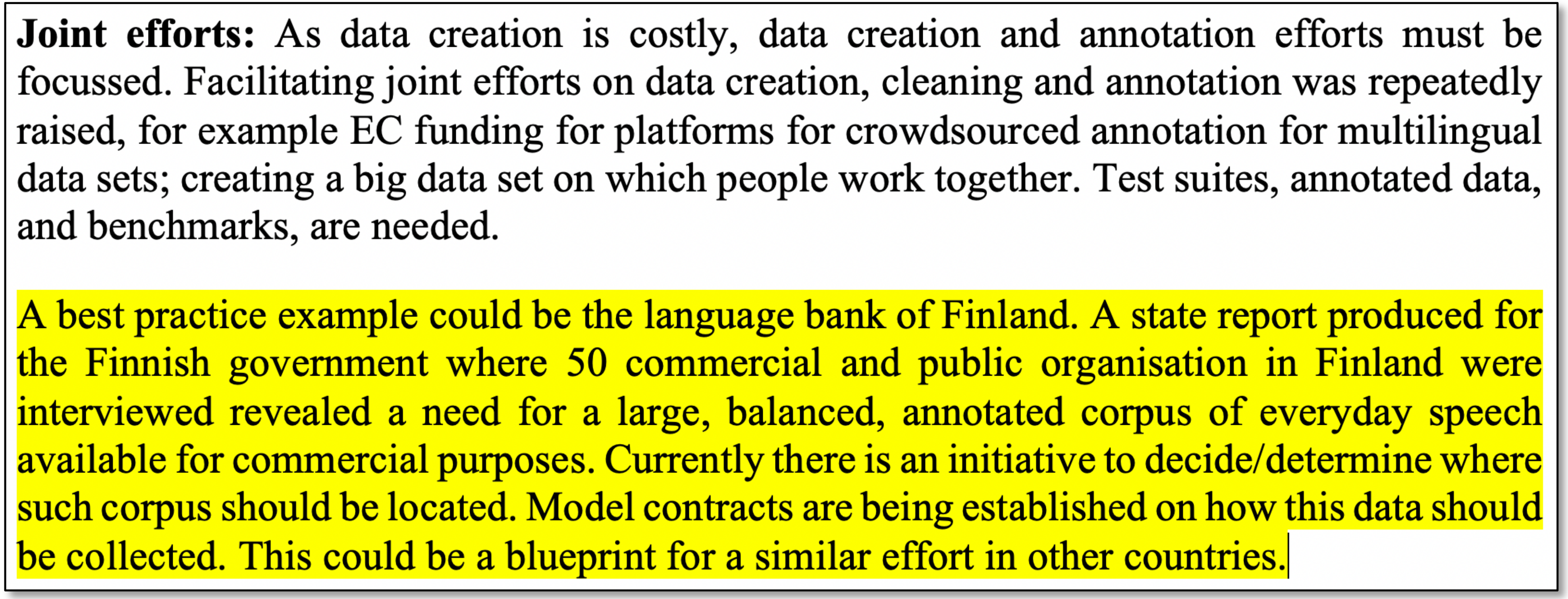
Marko Turpeinen
1001 Lakes

Background

Vake has actively supported the development of language-centric artificial intelligence solutions for Finnish languages (Finnish, Swedish and Sámi languages).

Progress:

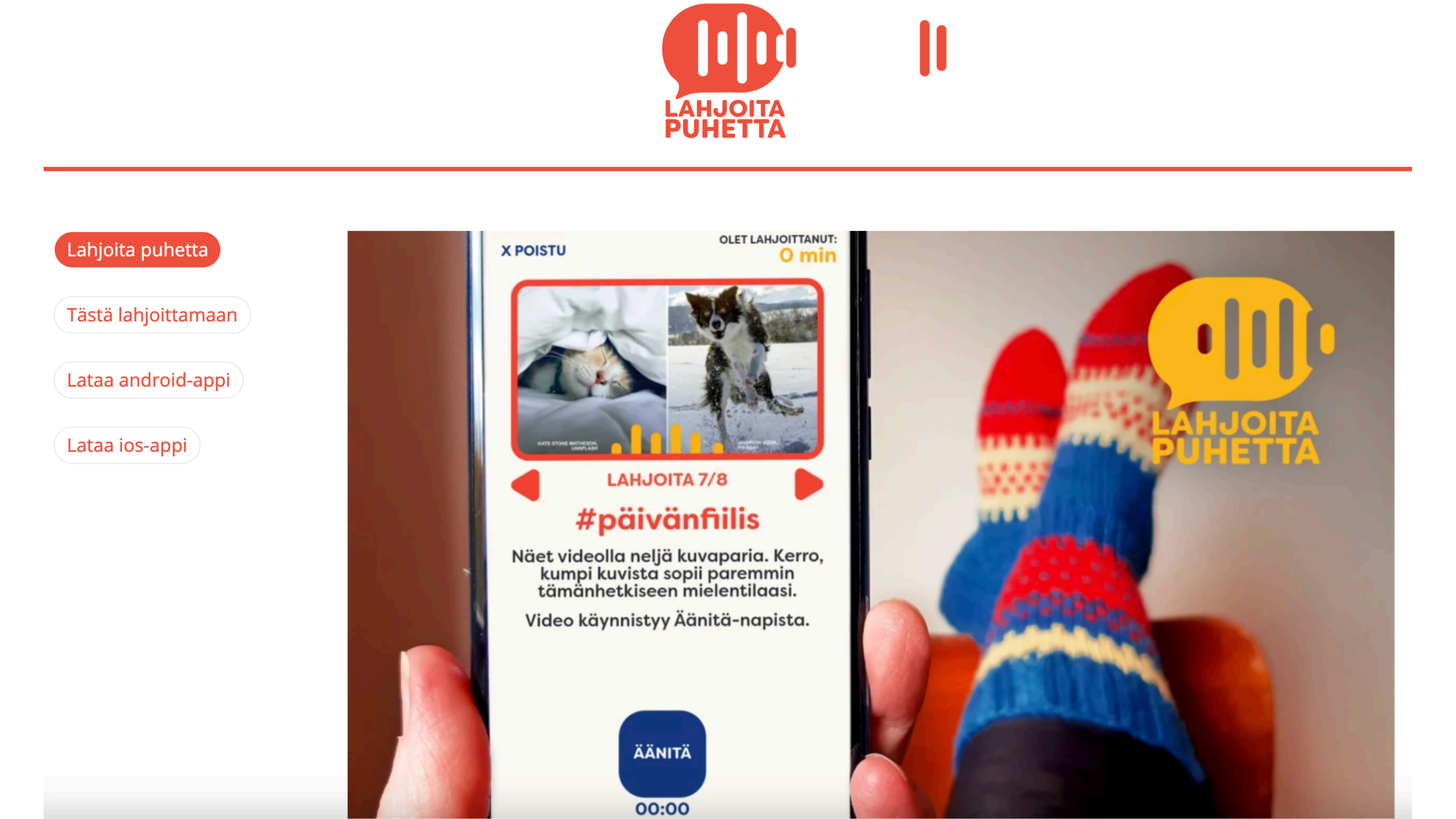
Q2-Q3/19	Pre-study
Q3/19-Q2/20	“Donate Speech” – campaign planning
Q4/19-Q1/20	“New Organization” – plan
Q4/19-Q2/20	Legal study
Q3/20-	“Donate Speech” – campaign running
Q3/20-	Entrepreneurial core group formed
Q4/20-	Investor discussion
Q4/20-	Business Finland project planning



Joint efforts: As data creation is costly, data creation and annotation efforts must be focussed. Facilitating joint efforts on data creation, cleaning and annotation was repeatedly raised, for example EC funding for platforms for crowdsourced annotation for multilingual data sets; creating a big data set on which people work together. Test suites, annotated data, and benchmarks, are needed.

A best practice example could be the language bank of Finland. A state report produced for the Finnish government where 50 commercial and public organisation in Finland were interviewed revealed a need for a large, balanced, annotated corpus of everyday speech available for commercial purposes. Currently there is an initiative to decide/determine where such corpus should be located. Model contracts are being established on how this data should be collected. This could be a blueprint for a similar effort in other countries.

Stakeholder Consultation on Language Technologies in Digital Europe, EU Commission - DG Connect, Brussels 10-11 February 2020 (draft of meeting report)



LAHJOITA PUHETTA

Lahjoita puhetta

Tästä lahjoittamaan

Lataa android-appi

Lataa ios-appi

X POISTU

OLET LAHJOITTANUT:
0 min

LAHJOITA 7/8

#päivänfiilis

Näet videolla neljä kuvaparia. Kerro, kumpi kuvista sopii paremmin tämänhetkiseen mielentilaasi.

Video käynnistyy Äänitä-napista.

ÄÄNITÄ

00:00

Societal and business need

Small language areas need to accelerate the use of their own languages in AI solutions

1. The importance of machine learning in services is growing

The importance of artificial intelligence in the services and business provided to citizens is growing. Finnish companies should be at the forefront of utilizing artificial intelligence.

One of the key growth areas of artificial intelligence is machine learning and especially its utilization in text and speech processing.

Utilizing machine learning to process text and speech in Finnish (national official languages and Sámi languages) requires models developed on the basis of large availability of training materials.

2. Small language areas lag behind in the development of AI

The field of language technology in Europe remains highly fragmented and the development of multilingual artificial intelligence has been a relatively minor side of wider EU initiatives to promote artificial intelligence.

Due to the relatively low level and lack of availability of language resources, artificial intelligence services for citizens and businesses are not developing fast enough for small languages.

The market and perceived market failure

There is no player in the market to accelerate the utilization of Finnish languages

1. The market for natural language applications is growing

The international market for language resources, natural language processing and AI applications based on these solutions is growing:

- new consumer services (chat interfaces, device voice control, voice-guided customer service)
- automatic translation solutions
- business analytics solutions based on text and speech recognition.

2. Small language areas are not international business priority

The most significant language models on the market are closed and primarily serve the business needs of the international companies that developed the models.

The priorities of international companies are in large markets (language areas), whereby solutions that use Finnish languages arrive only after larger language areas - if at all.

3. Finnish market is under-developed

The need for services based on the Finnish languages has been identified, but companies and public actors approach the problem independently and the resources are often aimed at research use only.

The most significant shortcoming is the large, easily accessible Finnish speech corpus and the language models of Finnish speech built with it.

In addition, scalable computing capacity is needed to complement scientifically oriented computing centres and international cloud services.

“LANGUAGE PLATFORM” (i.e. Kielialusta)

The Finnish market needs a new player that builds and manages models for implementing Finnish-language (national and Sámi) solutions that utilize machine learning. Due to the required investments, it is assumed that no new player will be created by itself, which will expand, activate and strengthen this network of actors.

Launched with the support of Vake, the “Language Platform” has two tasks:

Task 1: To create a cost-effective platform for basic language resources that fills gaps in the market and enables Finnish language language solutions.
Task 2: To increase the competitiveness of language-based artificial intelligence made in Finland and make it an internationally successful export product.

Language platform ecosystem

The language platform serves the entire Finnish language technology ecosystem

The value promise of the language platform for the ecosystem

1. Comprehensive, up-to-date and high-quality language resources that enable customers to make better artificial intelligence applications.
2. Address the market failure related to the underdevelopment of basic small language services, which prevents companies from developing their own services.
3. Industry players and language resources find each other better.
4. Operators with less language technology skills will also have easy access to skills related to advanced models.
5. Better understanding of the usefulness of language resources and market needs based on own statistics and analysis.
6. International marketing of Finnish ecosystem know-how and raising public awareness, development of operational processes in the field, EU influence.

How do companies and other organizations benefit from the Language Platform?

- availability and visibility of services
- market opening and faster market access
- cost savings in experimenting, developing and utilizing language resources
- finding language technology and artificial intelligence skills and partners

How does research benefit the Language Platform?

- conducting experiments and prototyping more efficiently and faster
- finding partners
- wider distribution and utilization of language resources in industry

Customers, products, incomes

High quality and comprehensive language-centric artificial intelligence resources

Customer segments

Public service providers

Private service providers

Developers of language technology and AI

Software developers

Research community

Services

Key activities

- Manages language materials, software and model libraries
- Process collected language materials (cleaning, metadata, transcripts)
- Create language templates
- Manage rights and licensing
- Acts as a matchmaker between language resource providers and users

Other functions

- Providing computing capacity
- Consulting
- Education and research
- Marketing of national know - how
- Statistics on the use of language resources

Potential income streams

Language technologies as functional services

- License fees
- API-based pricing
- Fee for the use of computing resources

The marketplace for language resources

- License fees
- Brokerage fee, matchmaking fee

Professional Services

- Consulting fees

Other

- Advertising and sponsorship
- Training
- Project financing
- Membership fees

Operative model for Language Platform

There are three main alternatives for operative model

Kielialusta

Expectations:

- Serves the needs of both the public sector and the commercial sector
- Constantly evolving based on market needs
- A de-facto player in the market but does not hinder or slow down the development of the commercial market

Public platform	+ Public sector savings from centralized operations + Enables easier movement of public data + De-facto player in Finland for public sector and EU cooperation - Does it adequately serve commercial needs? - Is fast enough cf. market development?
PPP (Public-Private-Partnership)	+ Serves the needs of both the public and commercial sectors + De-facto player in Finland for public sector and EU cooperation + Enables changes in the ownership base as the market evolves + Can operate on a non-profit principle - Is fast enough cf. market development?
For-profit company	+ The needs of the commercial sector are strongly involved + The most flexible ownership base as the market develops + Rapid development of operations - Will the company become a de-facto market player in terms of public sector and EU cooperation? - Does it serve the public sector in a sufficiently balanced way?

Next steps

1. **Leader.** Language Platform startup phase needs a dedicated leader person.
2. **Language Platform (PPP) participants and investors.** Specify the potential participants in the Language Platform and the planned development of the ownership base.
3. **Kielipankki (”Language Bank”).** Decide on the relationship between the Language Platform and Kielipankki, e.g. material acquisition, computing resources and expert resources.
4. **European Language Grid.** Plan the relationship between the Language Platform and ELG, avoiding duplication. The functioning of the Language Platform as a national ELG pilot will be investigated.
5. **Market Testing.** A more detailed analysis of the demand for the planned services of the Language Platform.
6. **Technology.** Define key technology options for the Language Platform solution. Description of the top-level technical architecture. Link with ELG.
7. **Agreements.** Prepare model contracts for the acquisition and use of language resources..

Founding the Legal Entity

Q1/2021

Thank you!

Marko Turpeinen
1001 Lakes
marko.turpeinen@1001lakes.com