

Increasing the impact of OH with HLT

How can advanced HLT help to increase the findability and (re-)usability of OH-collections in scholarly work?



I'm a computer scientist, Michael. That means that I can't explain about people, their motivation or their dreams. But if you want to know how to search for "words", how to combine "themes" or how to "access recordings", I'm your man.

Making sense of digital spaghetti

CLARIAH and Utrecht University are putting together a two-week immersion course on getting to grips with Big Data.

Location: Utrecht & Woudenberg

Date: 4-jan-2016 - 15-jan-2016

Time: 9:00 (welcome with coffee)

Background

Increasingly, research in all disciplines, from the natural sciences to social sciences and humanities, involves big data. The availability of vast amounts of textual, audio-visual and structured data from digital sources is revolutionizing research in the humanities and social sciences. The most advanced scholarship in these areas, currently and in the foreseeable future, relies on the use of sophisticated tools for accessing, processing, analysing and presenting this data.

General goals

CLARIAH, together with University College Utrecht, offer a short two-week module in which we give a small group of undergraduate, graduate students, and professional researchers the opportunity to gain familiarity and experience with some common approaches to handling very large datasets. Together, we'll use a batch of current Twitter™ data as a vehicle for practicing computational thinking, and the general concepts that data analysis with computers involves.

As part of an inclusive approach to large-scale research, this module stimulates the kind of thinking that CLARIAH hopes to engender: basic programming techniques, the use of multiple paradigms to solve problems, drawing on reasoning, logic, analysis, hypothesis-testing, and formal problem-solving methods, enabling all researchers, regardless of discipline, to engage fully with their own research.



Invisibility of OH-collections

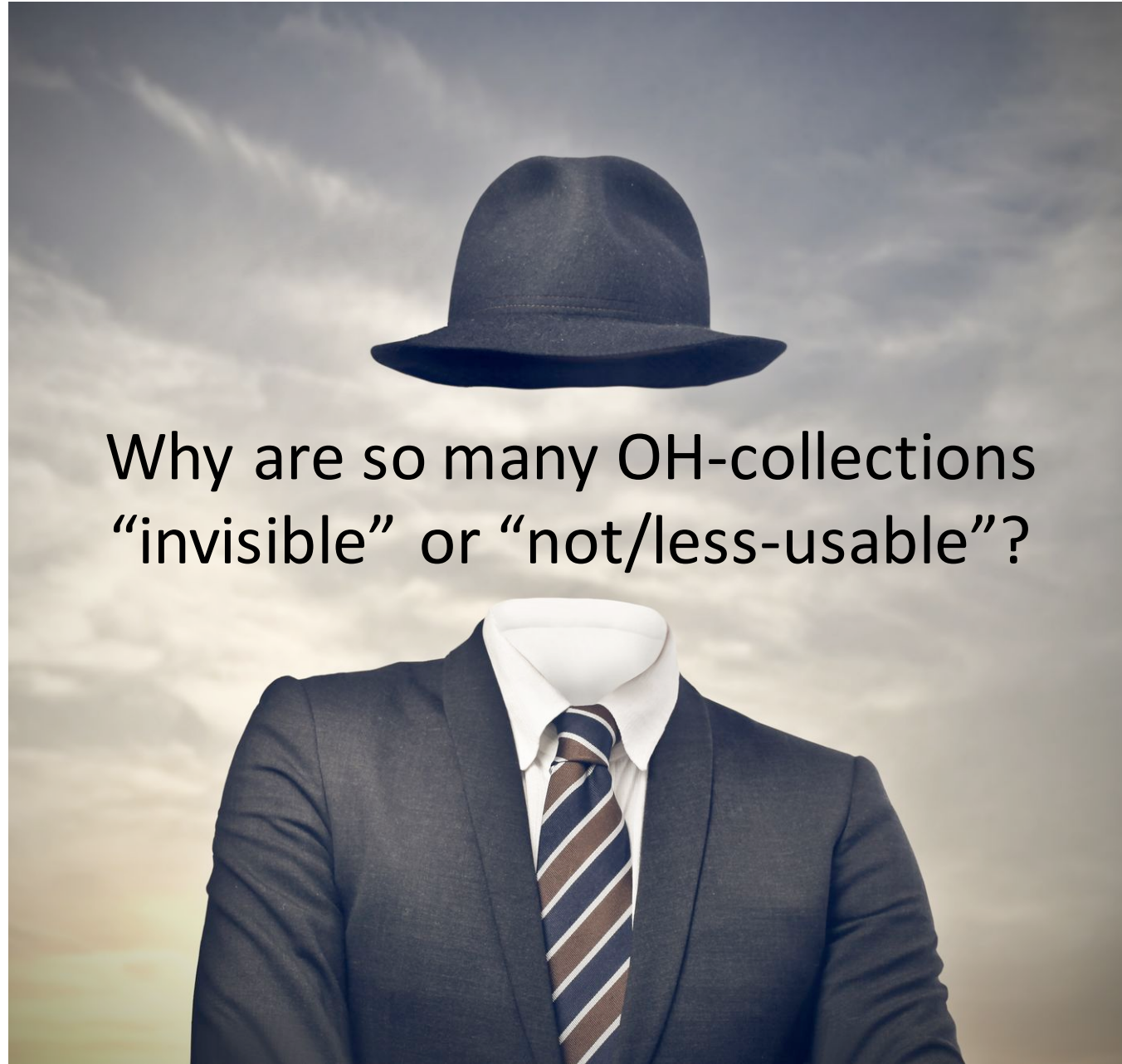
Lack of metadata

Non-open policy of the archives

IPR-rules

Not digitalized

Only in the local language



Why are so many OH-collections “invisible” or “not/less-usable”?

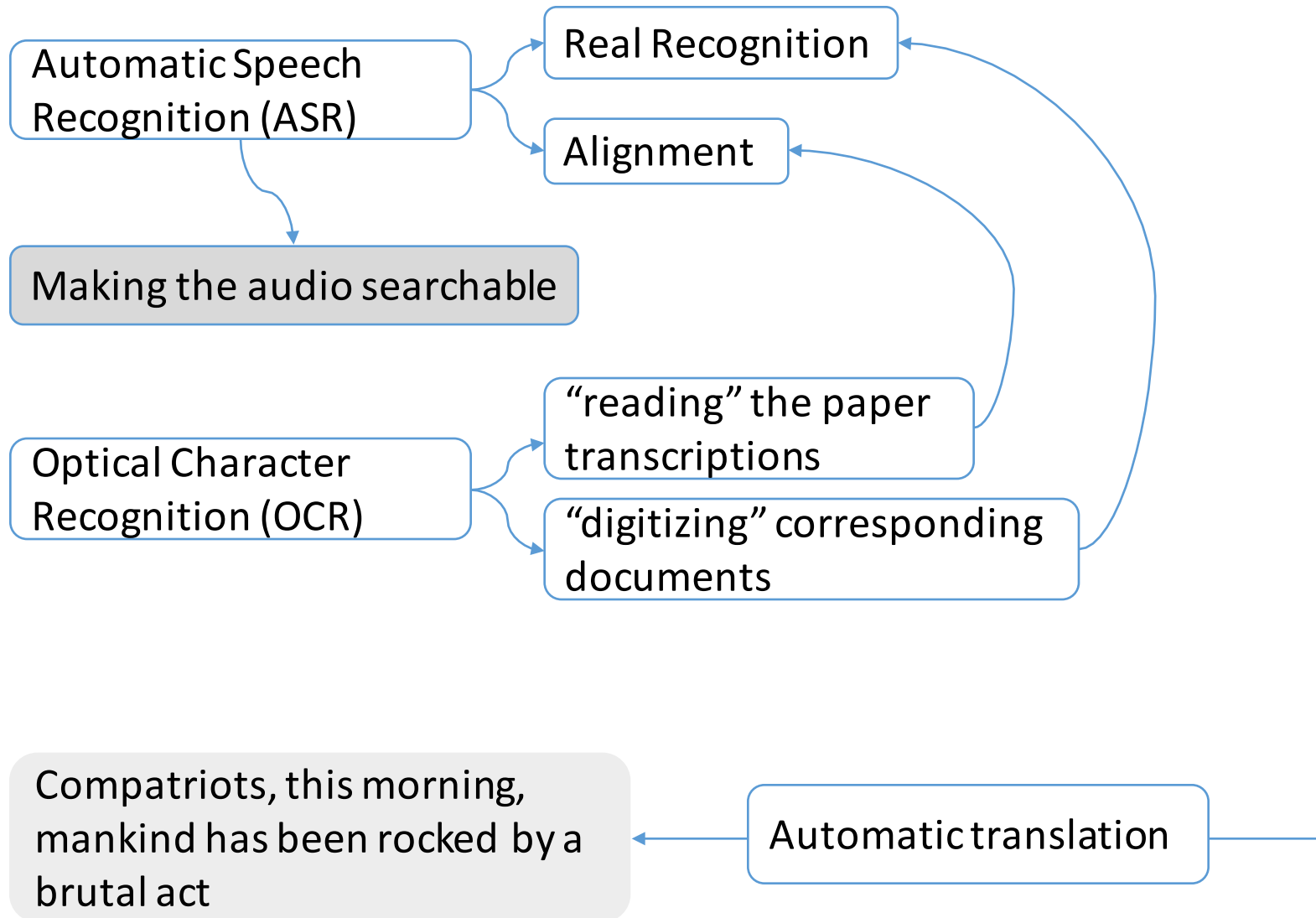
Lack of metadata standards

Lack of of fine-grained descriptions

Lack of the full recordings (*often just edited snippets*)

Lack of context

Human Language Technology (HLT)



*Landgenoten, heden morgen is **den menscheid** opgeschrikt door een brute daad*

Spelling Checking

*Landgenoten, vanochtend is de **mensheid** opgeschrikt door een brute daad*

Speech Recognition

DL: Impact

Strong improvement due to:

- Massive availability of data (speech and text)
- Unlimited computing power (cloud computing)
- Deep Learning Algorithms

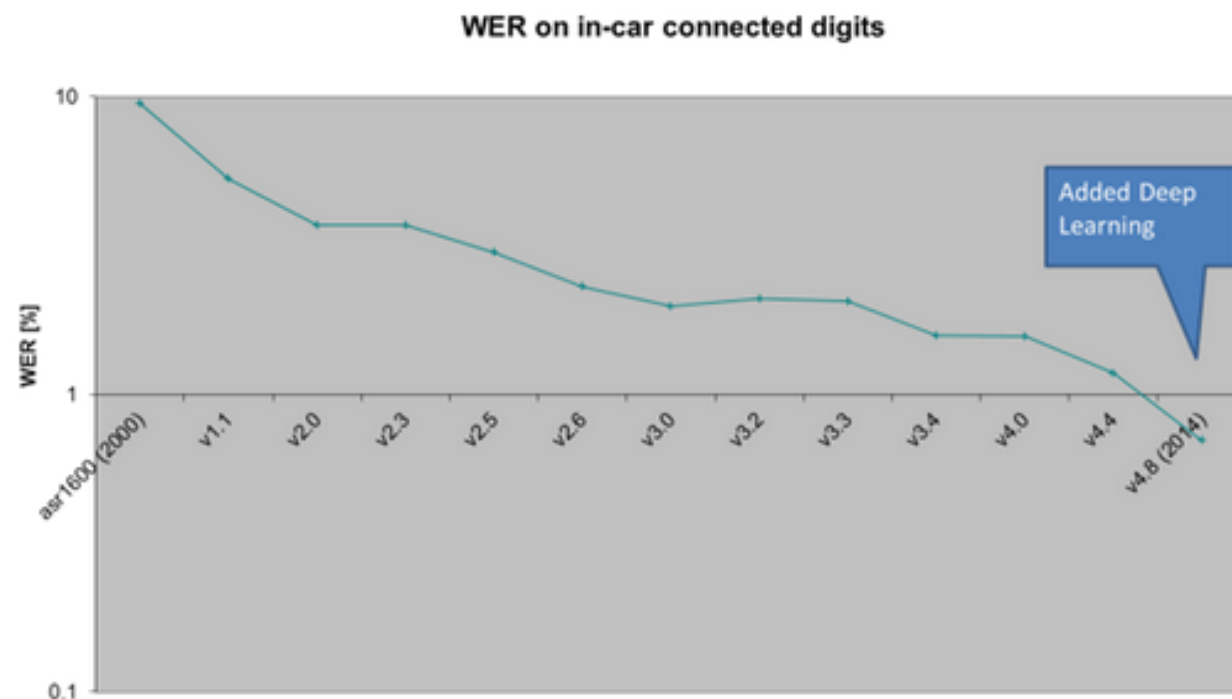
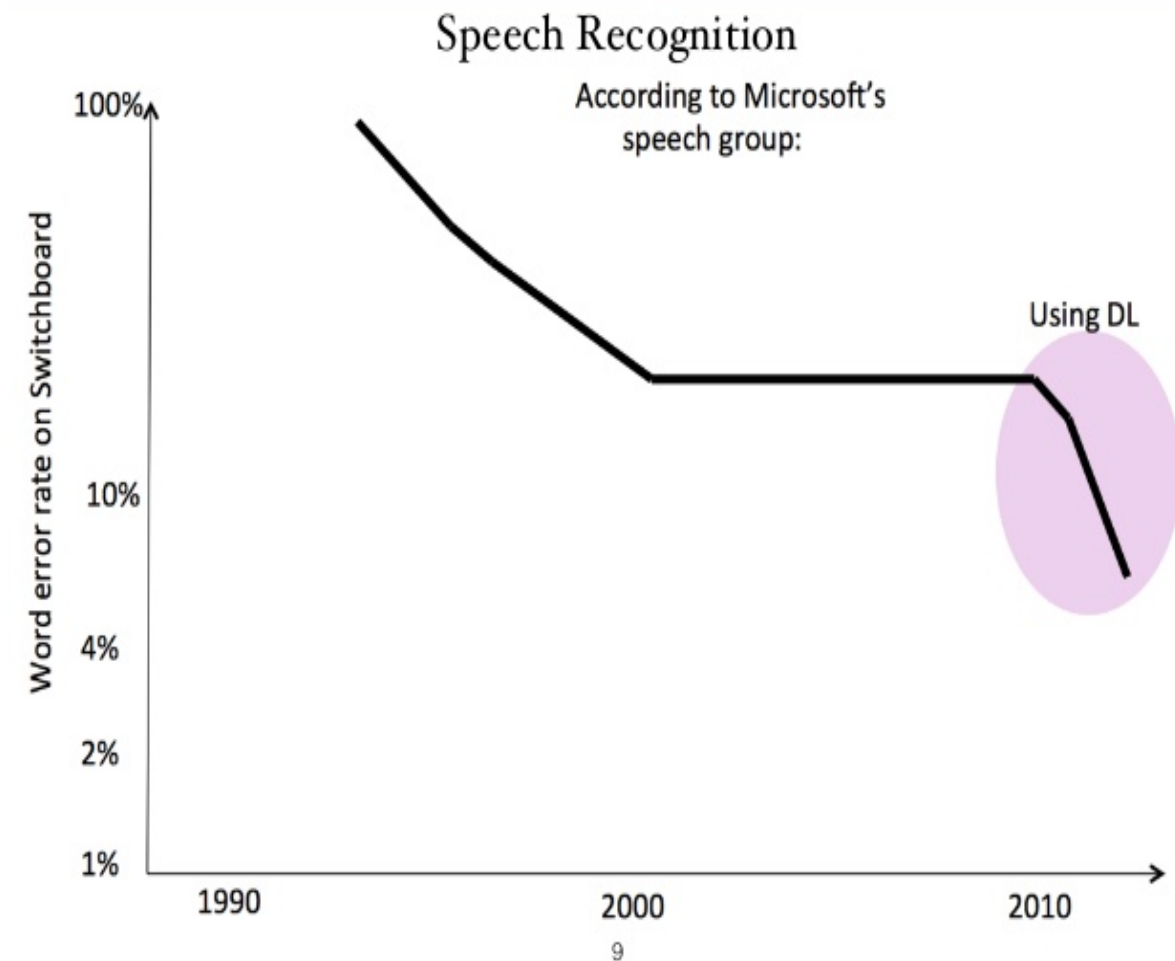
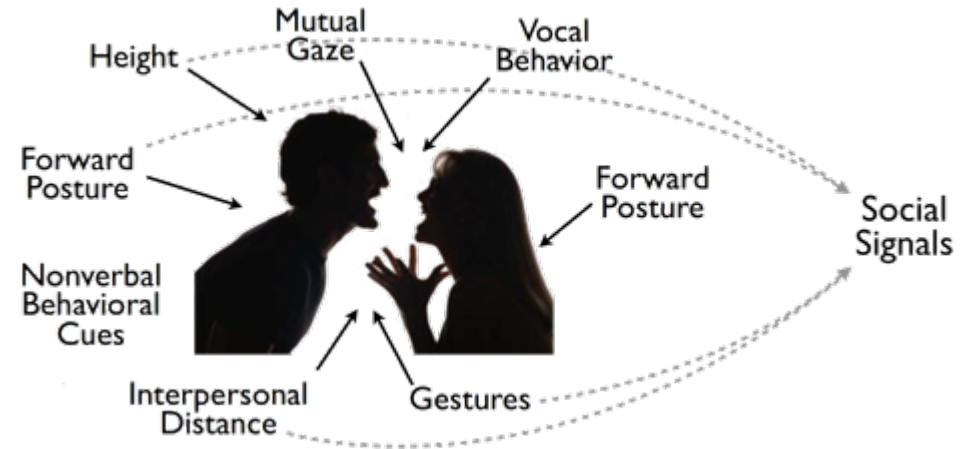
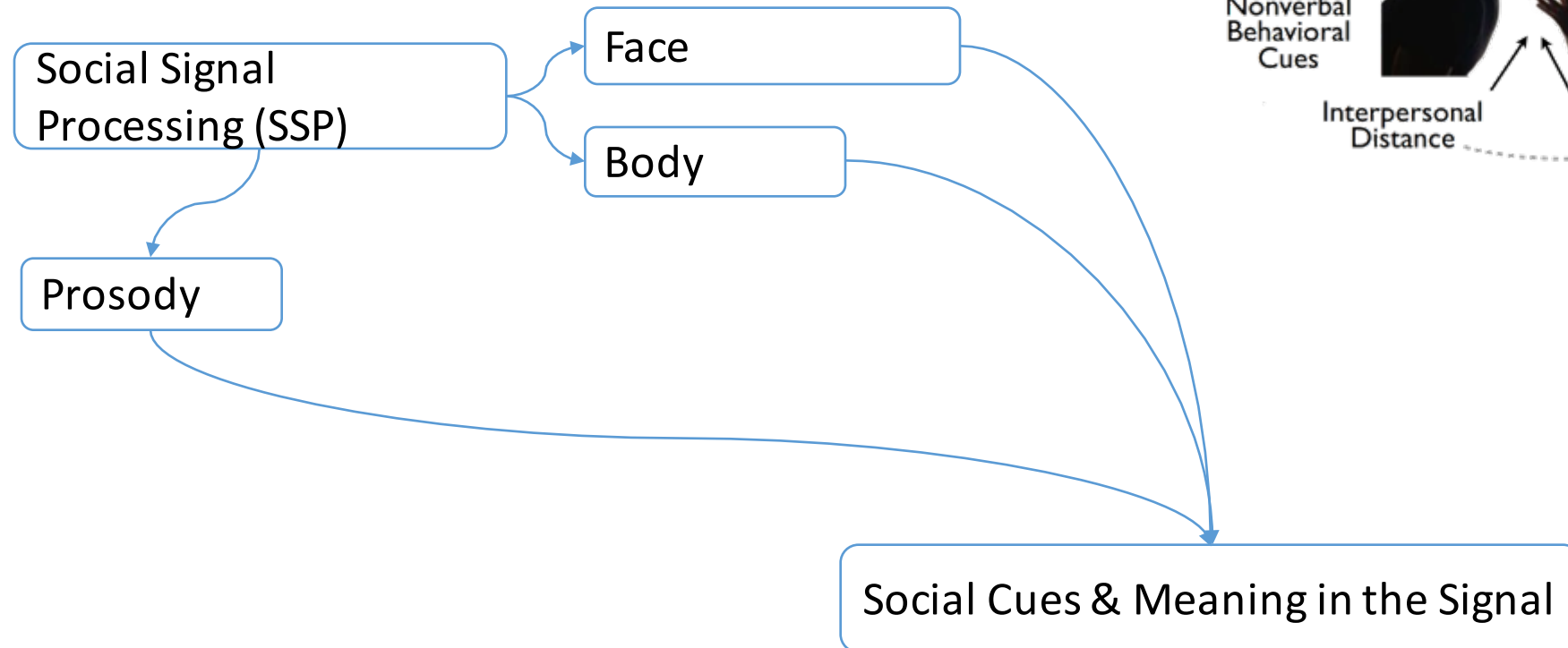


Figure 3: Exponential error rate reduction since 14 years

Human Language Technology (HLT)



What need to be done?



Facilitating the useful HLT-technology

Webservice to make/edit the metadata

Broadcasting the metadata of the OH-collections

Use WebVTT as the default subtitling format.



DynamicMetadata

Static Metadata

Creation of a flexible, comprehensive metadata scheme (CMDI?) with obligatory and potential fields.

WebVTT files provide captions or subtitles for video content, and also text video descriptions, chapters for content navigation, and more generally any form of metadata that is time-aligned with audio or video content.

Dynamic metadata



First 3 min



Middle 3 min



Last 3 min

TIME



Metadata

Static Metadata

File level

File level

File level

File level

File level

Collection level

What are they talking about?
Who are talking?
Which language(s)?
How can I access it?

(T1 → T2)

Topic 2
(T3 → T4)

Topic 2
(T5 → T6)

Chap. 1
(T1* → T2*)

Chap. 2
(T3* → T4*)

Speaker 1
(T1'' → T2'')

Speaker 2
(T3'' → T4'')

Speaker 1
(T5'' → T6'')

Speaker 1&2
(T7'' → T8'')



?

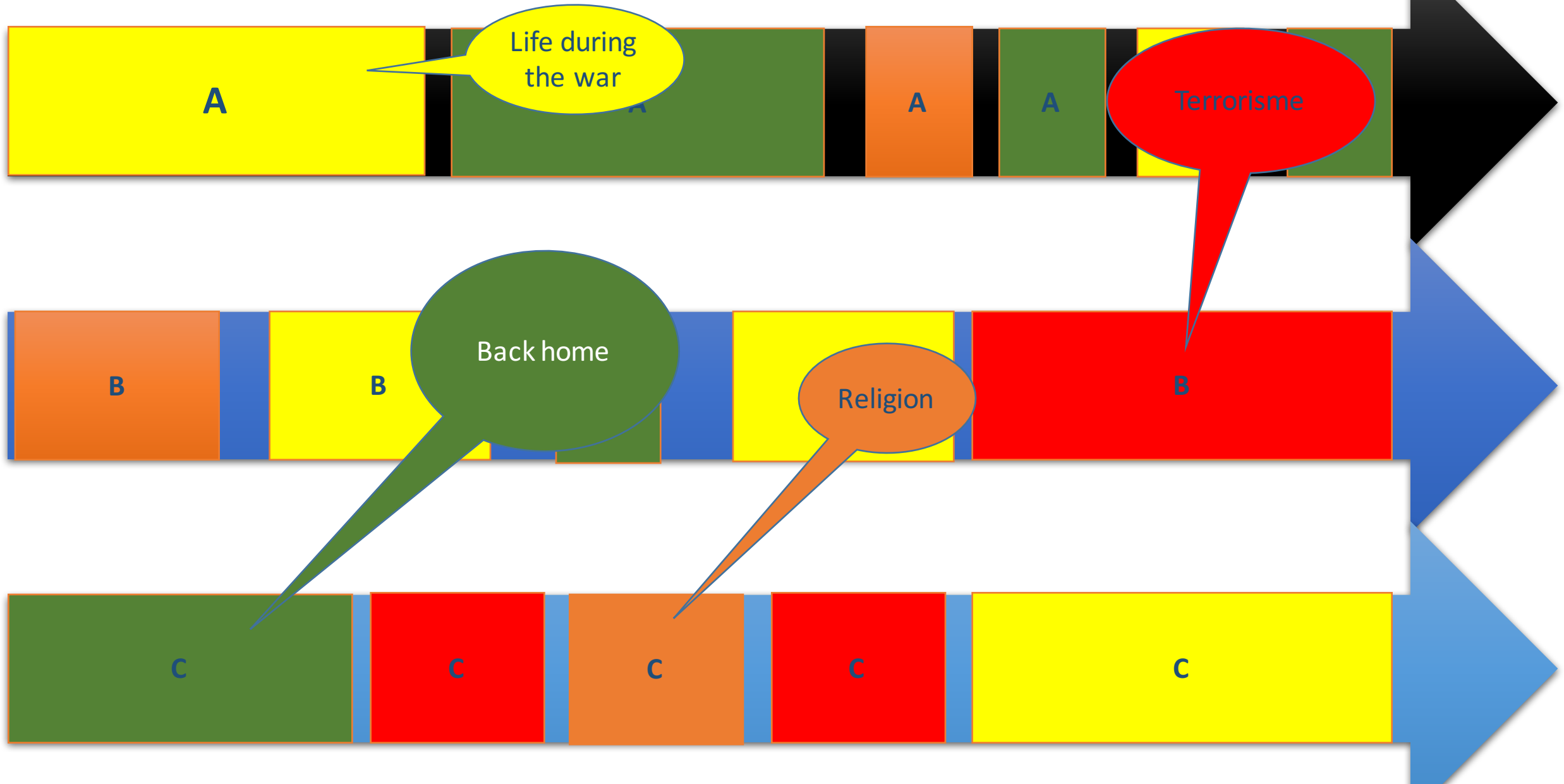
?

?

?

This was just one AV-file

Determine the topics-in-time of all your AV-files



Some Examples

Bosnian Memories

University of the Netherlands

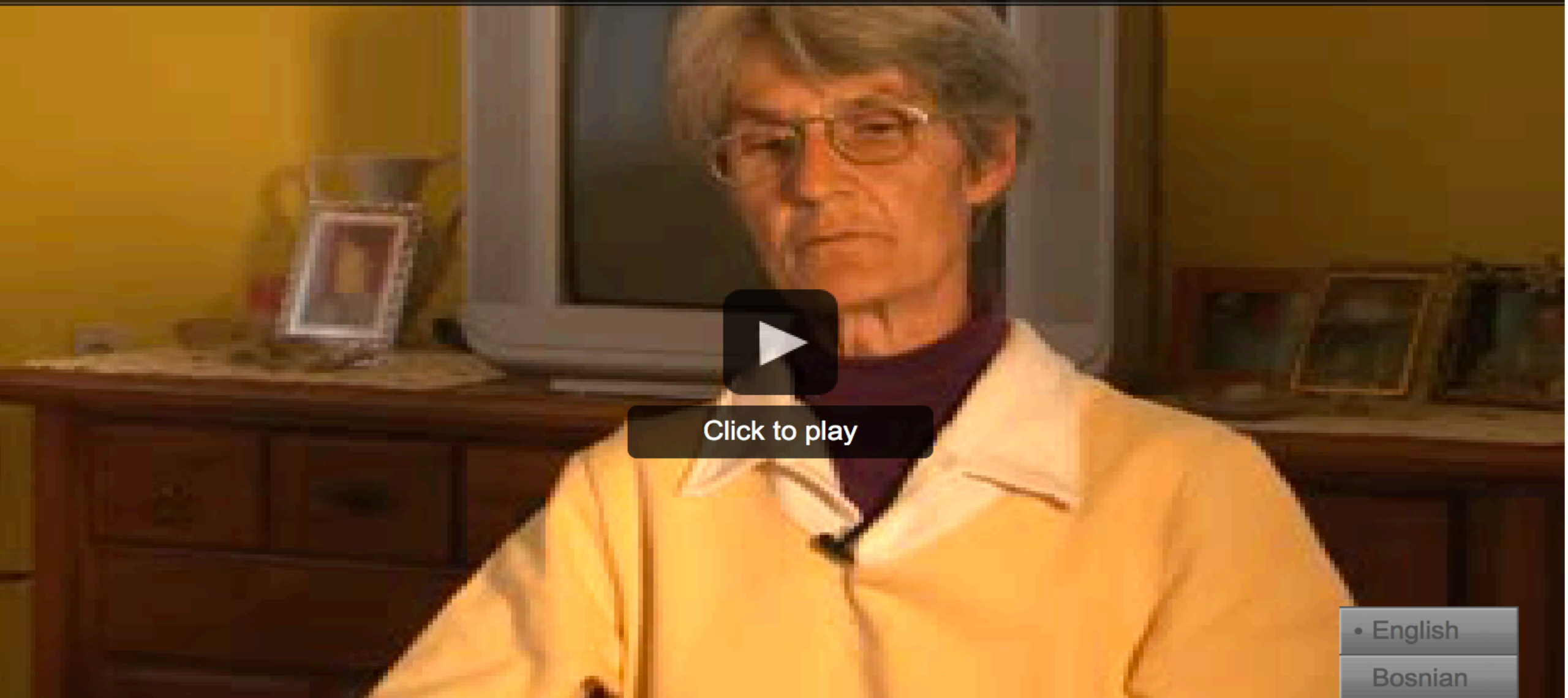
Back



Search bar

Blokkeren...

Andjelka Arnautalic
bihme



Click to play

- English
- Bosnian

Video player controls including play/pause, volume, progress bar (00:00 / 00:00), and resolution (720p HD).

University of the Netherlands



Een paar maanden later , toen de oorlog tegen Irak uitbrak . En toen ook nog in november 19 , 80 een

Increasing the impact



ASR

Imperfect result



Crowd sourcing

Perfect result

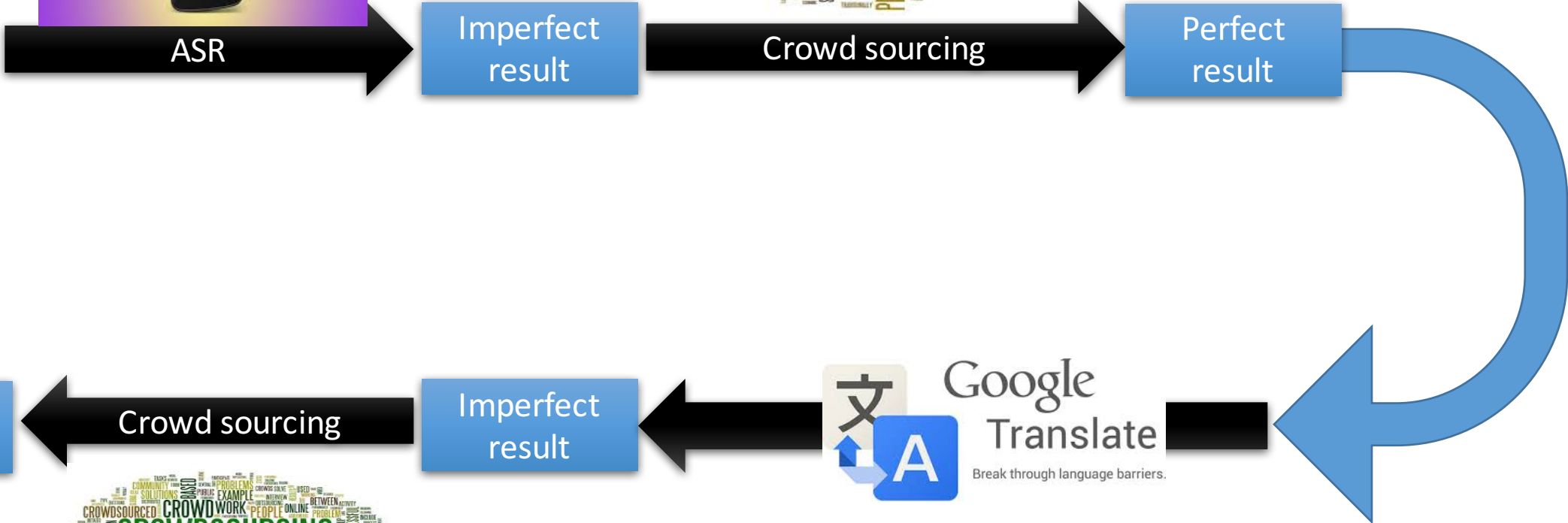
Perfect result

Crowd sourcing

Imperfect result



Google Translate
Break through language barriers.



Clustering software



The screenshot displays the Lingo3G Workbench interface, which is used for clustering documents. The main window is titled "Lingo3G Workbench" and has a menu bar with "File", "Search", "Window", and "About".

Search Panel: The "Source" is set to "Yahoo" and the "Algorithm" is "Lingo3G". Under the "Basic" section, the "Query" is "data mining" and the "Results" count is "400".

Clustering Results Panel: The title is "clustering (396 documents from Yahoo, 29 clusters from Lingo3G)". It shows a list of clusters and documents. The "Clusters" list includes: Clustering Algorithms (63), Clustering Methods (39), Server (40), Clustering Techniques (34), Cluster Analysis (31), Hierarchical Clustering (29), Document Clustering (26), Unsupervised Learning (14), Load Balancing (12), Solutions (17), High Availability (11), Search Results (10), Words (12), Microsoft Cluster (8), Example (10), Research (10), Open Source (7), Clustering Features (8), Topic (8), Network (8), and Clustering Support (7). The "Documents" list shows two documents: "[0] Cluster analysis - Wikipedia, the free encyclopedia" and "[1] Computer cluster - Wikipedia, the free encyclopedia".

Aduna Cluster Map Visualization: This panel shows a network graph with nodes and edges. Nodes are labeled with cluster names and counts: "Fuzzy Clustering (6/6)", "Document Clustering (1/2)", "Clustering Algorithms (2/63)", "Clustering Methods (1/39)", "Hierarchical Clustering (1/29)", "Clustering Techniques (34)", and "Spectral Clustering (5)".

Benchmark Panel: The status is "Round 497 (benchmark)". Performance metrics are: Avg time: 0.027 sec., Std dev: 0.003 sec., Min time: 0.020 sec., and Max time: 0.040 sec.

Attributes Panel: This panel shows settings for the "Minimum cluster size" (set to 0,0000) and "Minimum cluster size for subclusters" (set to 10). It also has checkboxes for "Normalize scores" (checked) and "Precise document assignment" (unchecked). There are expandable sections for "Debug", "Filters", "Global scorers", "Hierarchy", "Labels", "Language model", "Local scorers", and "Merging".

Attribute Info Panel: This panel provides details for the "Minimum cluster size" attribute. The "Description" states: "Determines the minimum allowed size of a cluster in relation to the parent cluster size. E.g. a value of 0.4 means that clusters must not contain less than 40% of the parent cluster's documents (of all documents in case of top-level clusters). This parameter is meaningful only if 'Document count label scorer weight' is greater than 0." The "Performance impact" is listed as "none".

System Tray: The bottom right corner shows "116M of 256M" memory usage and "Benchmarking... (88%)" progress.

Questions?