ON AIR:
HOW CAN TERMINOLOGY EXTRACTION AND MANAGEMENT TECHNOLOGY HELP LANGUAGE PROFESSIONALS IN BROADCAST MEDIA?

DANIYA KHAMIDULLINA

UNIVERSITY OF WOLVERHAMPTON
UNIVERSITY OF MALAGA

UNIVERSITY OF WOLVERHAMPTON

EMJMD in Technology for Translation and Interpreting

PRESENTATION STRUCTURE

Part I: Proposal of design of a digital tool for broadcast media linguists

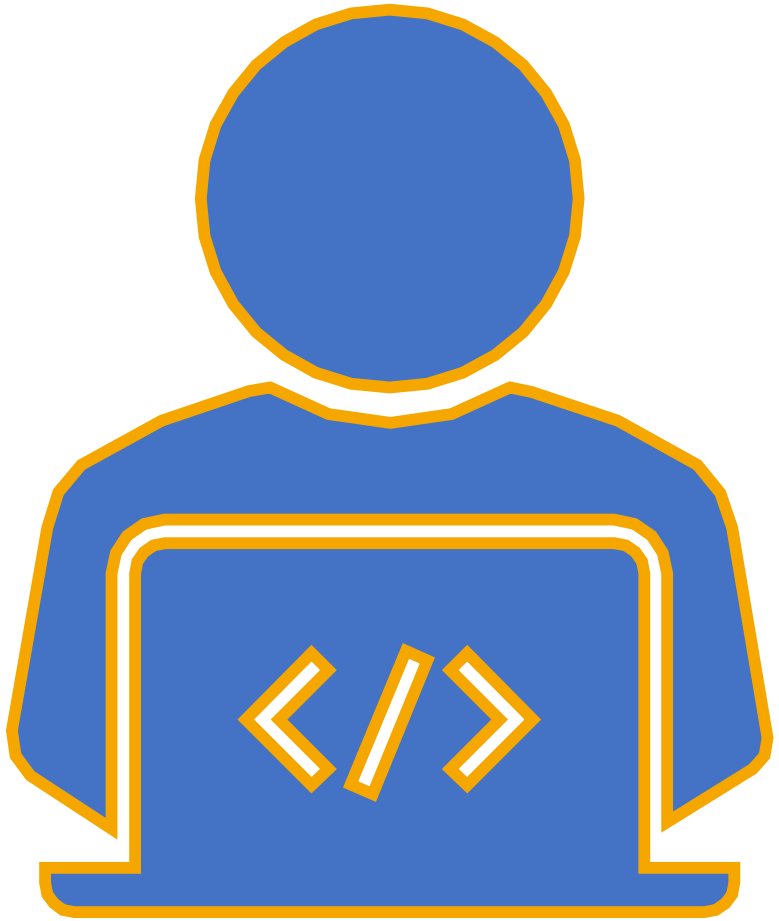
1. Project background
2. A tool for the media: in what ways should it be different from other CAT/CAI tools?
3. Key modules of the tool

Part II: Testing phase: Terminology extraction module (Russian-English)

1. Test dataset and methodology
2. OneClick Terms by Sketch Engine
3. Synchroterm by Terminotix
4. Conclusions
5. Future work

References

PART I: DIGITAL TOOL DESIGN PROPOSAL



1.1. PROJECT BACKGROUND

Moscow State University
(2010-2015)

Undergraduate thesis:
"Strategies For Rendering Information in Simultaneous Interpreting of Televised Interviews"
(Supervisor: Prof Andrei E. Levitsky)

University of Malaga
(2019-2020)

Interpreting Technology module term paper:
"Prototype of a CAI Tool for Broadcast Media Interpreters"
(Module Leader: Prof Gloria Corpas,
Lecturer: Mr Josh Goldsmith)

1.2. A TOOL FOR THE MEDIA: IN WHAT WAYS SHOULD IT BE DIFFERENT?

CONTEXT-SPECIFIC NEED

Harness available multilingual data (articles, transcripts, etc.)

SOLUTION

Built-in/interfaced web scraper

1.2. A TOOL FOR THE MEDIA: IN WHAT WAYS SHOULD IT BE DIFFERENT?

CONTEXT-SPECIFIC NEED

Ensure terminological consistency across platforms (e.g. website, radio, TV, DV) and regions

SOLUTION

Multimodal linguistic asset management solutions that cover different language combinations

1.2. A TOOL FOR THE MEDIA: IN WHAT WAYS SHOULD IT BE DIFFERENT?

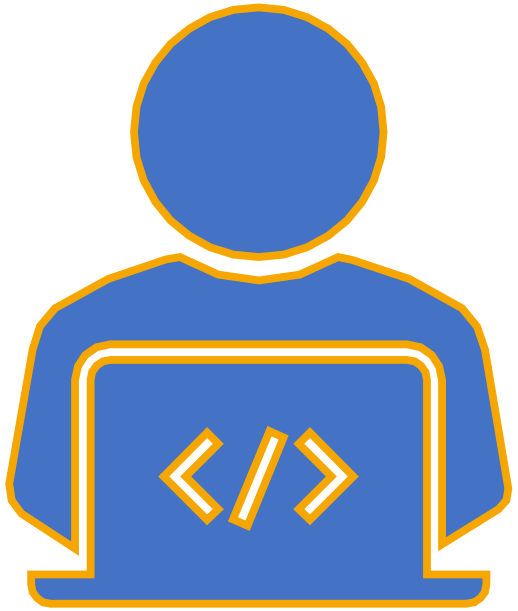
CONTEXT-SPECIFIC NEED

Increase marketability and practicality of the tool in the context of mass media

SOLUTION

Recyclable output – lines between different news production tasks are blurred (Bielsa, 2007, p. 143) so linguistic assets should ideally be transferrable

1.3. KEY MODULES OF THE TOOL



Terminology Extraction Module

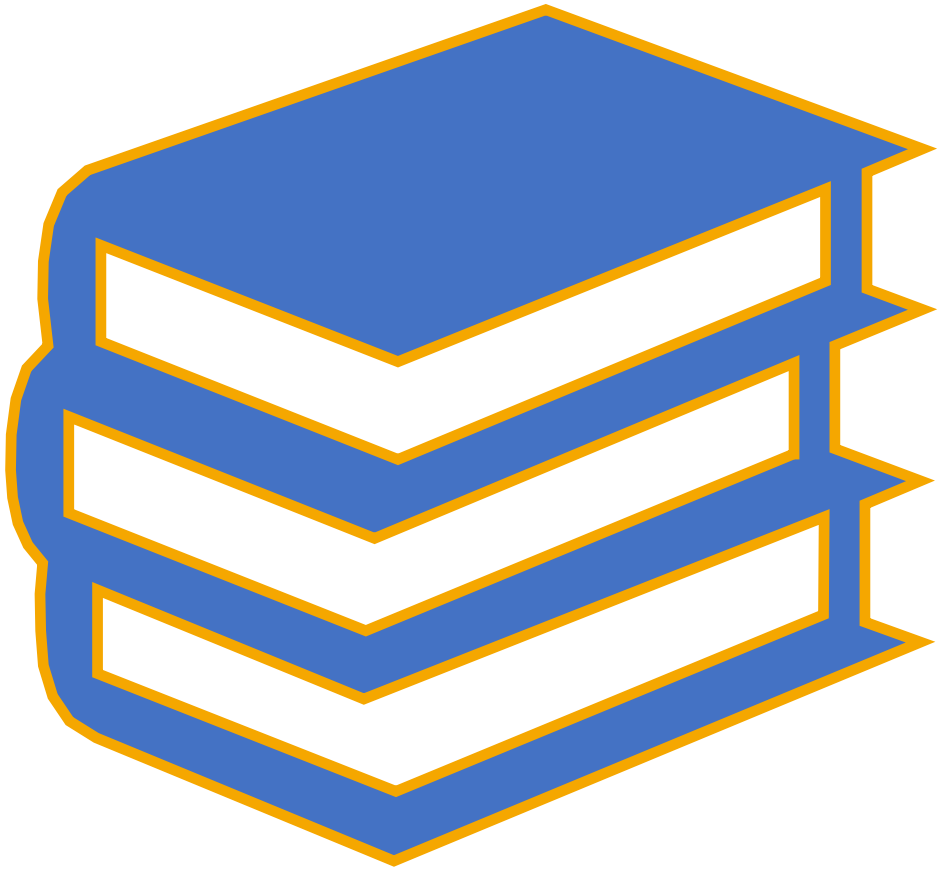
Terminology Management Module

Automatic Speech Recognition Module

1.3. KEY MODULES OF THE TOOL

Module	Stage	Relevance
Terminology Extraction	Assignment preparation	• automated term extraction **can increase terminological accuracy** during interpretation (Xu, 2018, p. 50) yet studies indicate that existing tools do not quite meet **the needs of interpreters** (Goldsmith, 2020, p. 299)
Terminology Management	Assignment preparation; Post-assignment debriefing; Adjacent language-related tasks	• a **collaborative approach** can help users **enhance term coverage and consistency** across domains (Costa, Corpas Pastor and Durán-Muñoz, 2018, p. 80) • can be used during onboarding of newly hired linguists
Speech Recognition	In the booth	• ASR (i.e. number, term or named entity recognition) could **improve interpreters' accuracy** as experimental studies have shown (Desmet, Vandierendonck and Defrancq, 2018, p. 25)

PART II: TERMINOLOGY EXTRACTION MODULE TESTING



2.1. TEST DATASET AND METHODOLOGY

- **Test situation**: interpreting a news conference from Russian into English
- **Dataset**: 10 publicly available transcripts of Vladimir Putin's annual news conferences (5 texts in Russian and 5 respective translations into English)
  - Downloaded in plain text format using an ad-hoc solution
  - Pre-processed manually (time and date information as well as tags removed)
  - Arranged into aligned bilingual transcripts with YOUALIGN
  - Result: **a parallel corpus of 267.898 words**

2.1. TEST DATASET AND METHODOLOGY

Ad-hoc plain text pulling solution

2.1. TEST DATASET AND METHODOLOGY

Two solutions tested:

Two ways of working with the dataset:

- extracting terminology from complete transcripts
- extracting terminology from thematic subcorpora created from these transcripts

2.1. TEST DATASET AND METHODOLOGY

MANUALLY CREATED THEMATIC SUBCORPORA

1. Agriculture and Aquaculture
2. Defence
3. Domestic Politics
4. Economy
5. Energy
6. Environmental Issues
7. Healthcare
8. Industry
9. International Relations – China
10. International Relations – Middle East
11. International Relations – Turkey
12. International Relations – Ukraine
13. International Relations – USA
14. International Relations – various
15. Social Affairs
16. Sports
17. Transport

2.1. TEST DATASET AND METHODOLOGY

MANUALLY CREATED THEMATIC SUBCORPORA

1. Agriculture and Aquaculture
2. Defence
3. Domestic Politics
4. Economy
5. Energy
6. Environmental Issues
7. Healthcare
8. Industry
9. International Relations – China
10. International Relations – Middle East
11. International Relations – Turkey
12. International Relations – Ukraine
13. International Relations – USA
14. International Relations – various
15. Social Affairs
16. Sports
17. Transport

2.2. *ONECLICK TERMS* BY SKETCH ENGINE (EXTRACTION IN BULK)

- **Extraction mode**:
  - Automatic monolingual term and keyword extraction

- **Result**:
  - Thematically heterogeneous output -> unlikely to be useful as glossary basis
  - Top entries belong to general language
  - Entries in Russian often non-lemmatized

OneClick terms

SKETCH ENGINE

2.2. *ONECLICK TERMS* BY SKETCH ENGINE (EXTRACTION IN BULK)

Most MWEs belong to common language:

- *e.g.* top 10 entries for English: *good afternoon, news conference, first point, second question, retirement age, defence industry, Russian economy, first question, news agency, tv channel*

- *e.g.* top 15 entries for Russian: *добрейший день, соединенный штат, средств массовой информации, пенсионный возраст, следующим год, лучший показатель, экономический союз, уважаемый Владимир, центральный банк, Евразийский экономический союз*

OneClick terms

SKETCH ENGiNE

2.2. *ONECLICK TERMS* BY SKETCH ENGINE (EXTRACTION IN BULK)

Some entries in Russian were non-lemmatized:

- *e.g. средств массовой информации (mass media) – genitive case*

And some contained declension conflicts:

- *e.g. следующим год (next year) – instrumental case + nominative/accusative case*

2.2. *ONECLICK TERMS* BY SKETCH ENGINE (THEMATIC SUBCORPORA)

- **Extraction mode**:
  - Automatic monolingual term and keyword extraction

- **Result**:
  - Output more thematically homogeneous than in bulk setup
  - More entries belong to specialized language
  - Some entries in Russian still non-lemmatized
  - Cases of possible source text misprocessing (e.g. grammatical gender swap)

OneClick terms

SKETCH ENGINE

2.2. *ONECLICK TERMS* BY SKETCH ENGINE (THEMATIC SUBCORPORA)

Most MWEs belong to specialized language:

- *e.g.* **Healthcare** domain top 10 entries for English: *medical assistance, primary care, cancer treatment, (own) pharmaceutical industry, medical air service, system-wide solution, head doctor, medical air, air service, child mortality*

- *e.g.* **Healthcare** domain top 10 entries for Russian: *тариф омс, первичное звено, уровень заработной платы, лекарственный препарат, следующим год, данные минфина, строительство онкоцентров, рядовой врач, системное решение, звено здравоохранения*

2.2. *ONECLICK TERMS* BY SKETCH ENGINE (THEMATIC SUBCORPORA)

Yet some entries in Russian were non-lemmatized:

- *e.g. ростом экономики (lit. 'by economic growth') – instrumental case*

Some also contained declension conflicts:

- *e.g. многополярного мир (lit. 'of a multipolar world') – genitive case + nominative/accusative case*

And some were repetitive (term boundary problem?):

- *e.g. natural population growth, natural population*

OneClick terms

SKETCH ENGiNE

2.3. *SYNCHROTERM* BY TERMINOTIX (EXTRACTION IN BULK)

- **Extraction mode**:
  - Automatic bilingual term extraction, batch processing

- **Result**:
  - Thematically heterogeneous output -> unlikely to be useful as glossary basis
  - In some cases, entry alignment is somewhat off
  - Some entries are non-lemmatized

SYNCHROTERM
TERMINOTIX

2.3. *SYNCHROTERM* BY TERMINOTIX (EXTRACTION IN BULK)

Output is predictably thematically heterogeneous:

Source Entry life news
Target Entry life news

Source Entry бюджетной сфере
Target Entry public sector

Source Entry ветеранов и инвалидов
Target Entry including disabled war veterans

Source Entry внутренний спрос
Target Entry domestic demand

Source Entry военной разведки
Target Entry military intelligence

Source Entry вчера только
Target Entry just yesterday

Source Entry газета «коммерсантъ
Target Entry kommersant newspaper

Source Entry говорит москва
Target Entry govorit moskva

Source Entry гоменюк-кравцова
Target Entry maria gomenyuk-kravtsova

Source Entry долларов за баррель
Target Entry barrel

A fragment of batch extraction output

SYNCHROTERM
TERMINOTIX

2.3. *SYNCHROTERM* BY TERMINOTIX (EXTRACTION IN BULK)

Some entries are partially misaligned:

- *e.g. долларов за баррель (lit. 'dollars per barrel') – barrel*

And some are non-lemmatized:

- *e.g. бюджетной сфере (public sector) – prepositional case*

SYNCHROTERM
TERMINOTIX

2.3. *SYNCHROTERM* BY TERMINOTIX (THEMATIC SUBCORPORA)

- **Extraction mode (I)**:
  - Automatic bilingual term extraction, batch processing
- **Result**:
  - Output contains noise and misalignments:

e.g. **Healthcare** domain: *алмазовский центр – addition to this hospital, внутрь самой отрасли – need to look, вообще не останется – change anything, впервые включена – put on that list, врачей – совсем другая – higher than ordinary doctors*

SYNCHROTERM
TERMINOTIX

2.3. *SYNCHROTERM* BY TERMINOTIX (THEMATIC SUBCORPORA)

- **Extraction mode (II)**:
  - Automatic bilingual term extraction, manual term selection and validation
  - More time-consuming but eliminates the need in mass PE

- **Result**:
  - A ready-to-use curated termbase

SYNCHROTERM
TERMINOTIX

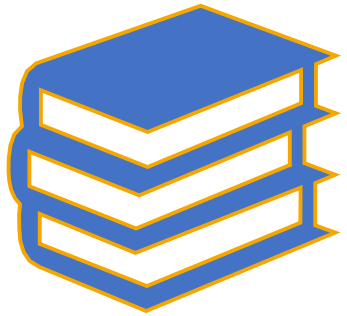
2.3. *SYNCHROTERM* BY TERMINOTIX (THEMATIC SUBCORPORA)

1	ВВП	GDP
2	Евразийский экономический союз	Eurasian Economic Union
3	НДФЛ	personal income tax
4	Резервный фонд	reserve fund
5	Фонд национального благосостояния	National Welfare Fund
6	Центральный банк	Central Bank
7	высокотехнологичные сферы	high-tech industries
8	дефицит бюджета	budget deficit
9	диспропорции на рынке	market disproportions
10	доходы населения	income of the population

A fragment of the Economy domain termbase, edited manually

SYNCHROTERM
TERMINOTIX

2.4. CONCLUSIONS



- Semi-automated generation of bilingual term lists from thematically arranged news conference subcorpora appears to yield output that requires the least amount of post-editing

- In the given scenario, it might be beneficial to enrich transcript-based thematic subcorpora with additional relevant materials to improve the quality of automated term extraction output

2.5. FUTURE WORK

- Further **terminology extraction tests** on **subcorpora enriched with additional thematic materials** could be run to see if that improves output quality
- **Corpus pre-processing could be automated** (e.g. tags/dates could be removed using scripts)
- To speed up detection of key topics and creation of thematic subcorpora, such NLP techniques as **topic analysis** could be employed
- Given that news conferences tend to be high-context events, it may be useful to **try generating domain-specific lists of named entities** using NER tools (e.g. *Natasha* (https://natasha.github.io/demo) for Russian)

REFERENCES

1. Bielsa, E. (2007) Translation in global news agencies, *Target*, **19**(1), pp. 135–155.
2. Corpas Pastor, G., Durán-Muñoz, I. and Costa, H. (2018) Assessing Terminology Management Systems for Interpreters, In *Trends in E-Tools and Resources for Translators and Interpreters*, Leiden, Brill, pp. 57–84.
3. Desmet, B., Vandierendonck, M. and Defrancq, B. (2018) Simultaneous interpretation of numbers and the impact of technological support, In *Interpreting and technology*, Fantinuoli, C. (ed.), Berlin, Language Science Press, pp. 13–27.
4. Goldsmith, J. (2020) Terminology extraction tools for interpreters, In *Interdependence and innovation in translation, interpreting and specialized communication*, Ahrens, B. (ed.), Frank & Timme: Verlag für wissenschaftliche Literatur, pp. 279–302.
5. Xu, R. (2018) Corpus-based terminological preparation for simultaneous interpreting, *Interpreting*, **20**(1), pp. 29–58.

ON AIR:
HOW CAN TERMINOLOGY EXTRACTION AND MANAGEMENT TECHNOLOGY HELP LANGUAGE PROFESSIONALS IN BROADCAST MEDIA?

DANIYA KHAMIDULLINA

UNIVERSITY OF WOLVERHAMPTON
UNIVERSITY OF MALAGA

UNIVERSITY OF WOLVERHAMPTON

EMJMD in Technology for Translation and Interpreting