

The logo for MORAVIA, featuring the word "MORAVIA" in a bold, red, sans-serif font. The letters are closely spaced and have a slight shadow effect. The background of the slide features several light blue, curved lines that sweep across the top and right sides, creating a sense of motion and flow.

MORAVIA

Flexible thinking. Reliable delivery.

Real-World Application of a Machine Translation Workflow

Tomáš Fulajtár

MORAVIA IT

Prague, September 7th, 2015

Personal Introduction



- 10+ years in Moravia
 - 8 years of engineering on SW localization (all OS platforms, QA work)
 - 2+ years in Language Technology Group (managing Moses environment)
 - Focus on MT system training, optimizations, technology improvements

Tomáš Fulajtár

tomasfu@moravia.com

+420 545 552 340

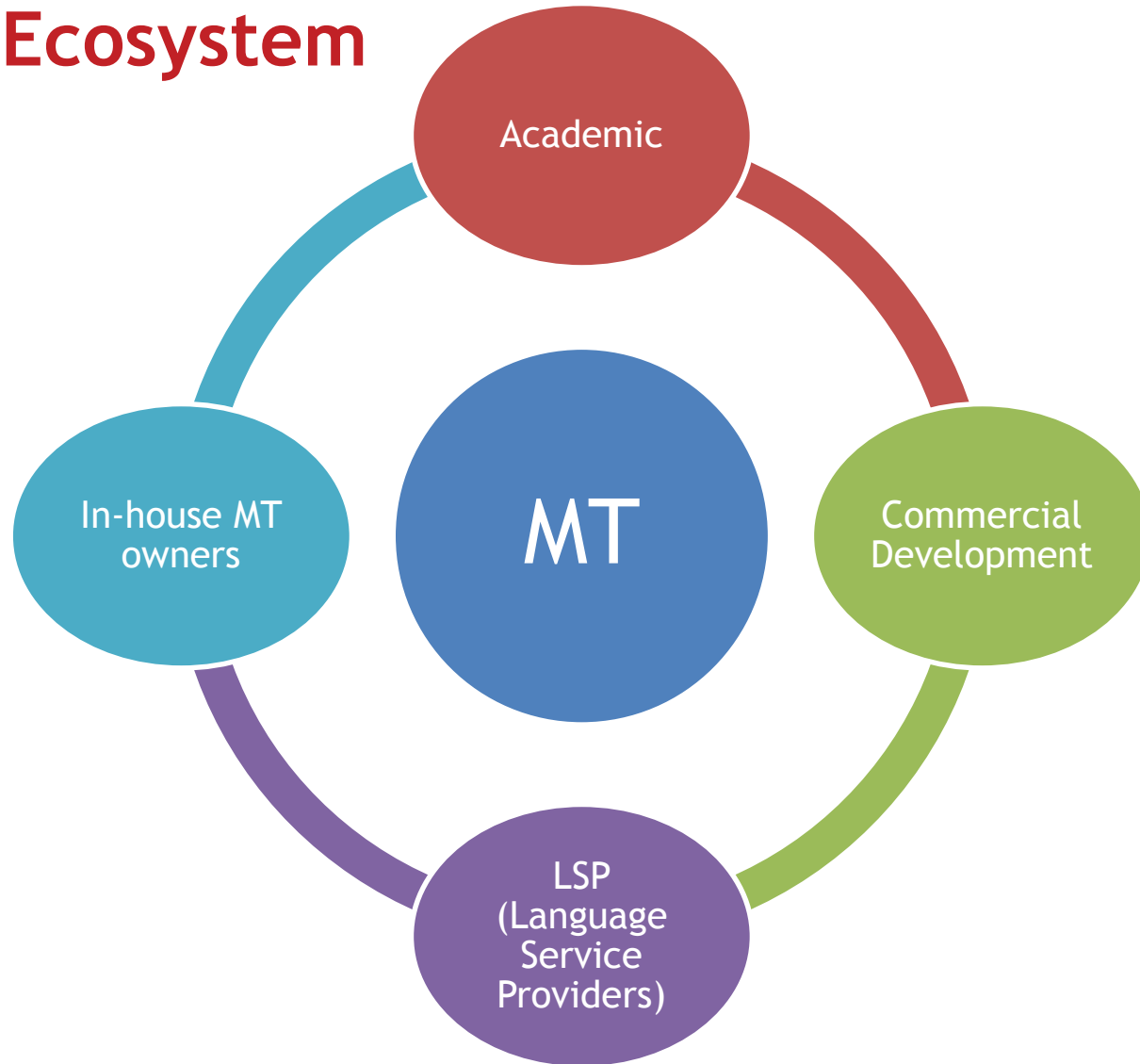
Overview

- The Machine Translation Ecosystem
- Localization 101
 - Translation Memory
 - Computer Aided Translation Tools
 - Combine TM + MT
- Challenges in applied MT
 - Moravia experience (solutions)
- Case Studies
- Open Topics
- Summary

The image features a solid red background. In the lower half, there are three overlapping, horizontally-oriented white ovals. The ovals are arranged in a way that they appear to be layered, with the top one slightly offset to the right and the bottom one slightly offset to the left. The text 'MT Ecosystem' is positioned in the lower-left area, partially overlapping the bottom-most oval.

MT Ecosystem

The MT Ecosystem



Language Service Providers

- Aka Translation Agencies
- Well established, yet dynamic industry
- Provide numerous linguistic services for clients
- Deal with
 - Variety of customers/products/content (IT, medical, automotive, etc.)
 - Huge number of languages (150+)
 - Wide range of CAT tools

Moravia at a Glance



Global
Offices: 10+

Internal
Headcount:
800+

Specialist
Translators:
5000+

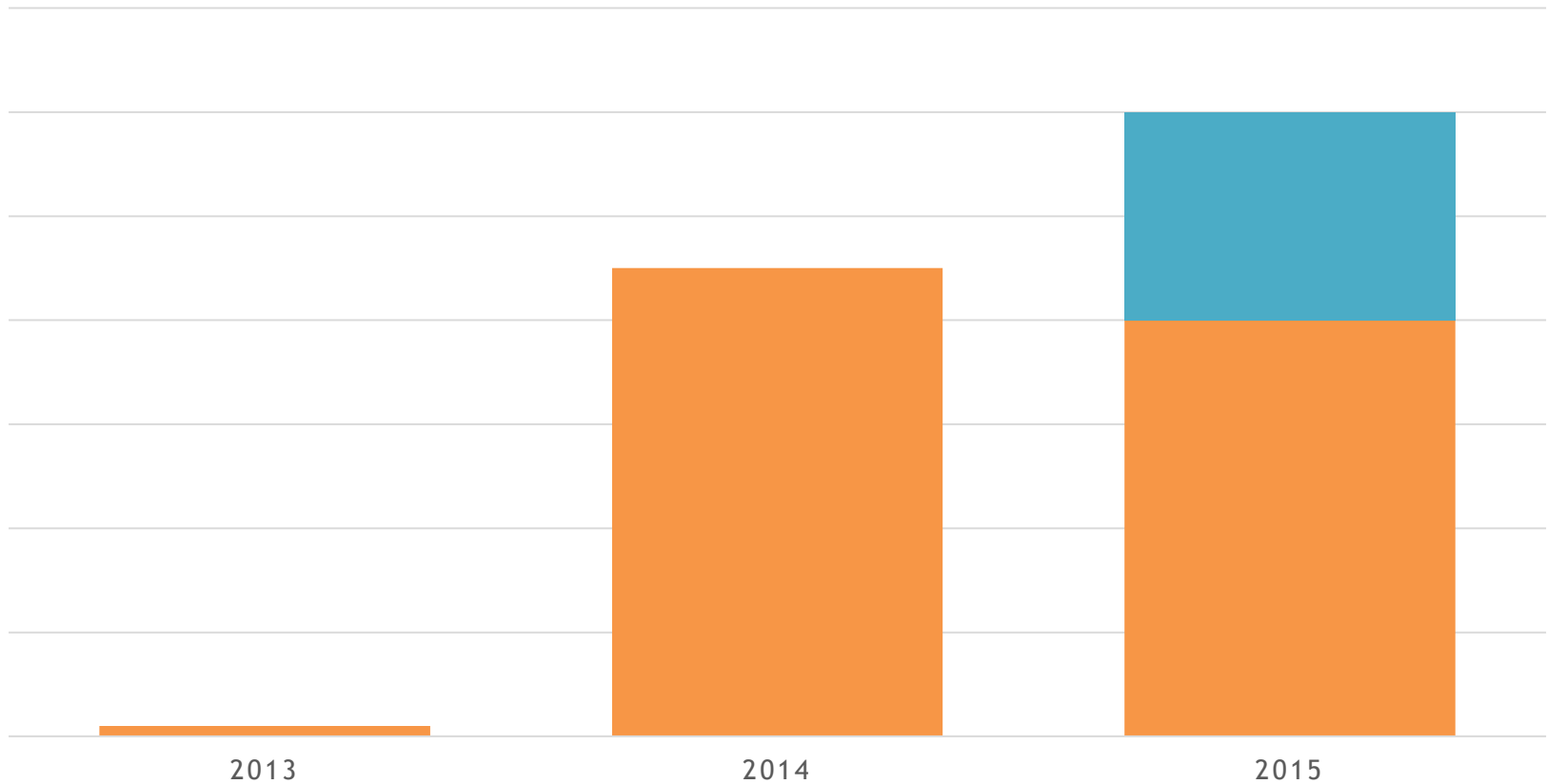
Languages:
120+

CAT tools:
12+

MT Engines:
100+

Growth of MT in production environment

■ MT ■ Prediction



The image features a solid red background. In the lower-left quadrant, the text "Localization 101" is written in a white, bold, sans-serif font. To the right of the text, there are three overlapping, white, curved lines that sweep from the left side towards the right, creating a sense of motion or a stylized graphic element.

Localization 101

Localization workflow 101

- Based on Translation Memory (TM) technology developed in the early 1990s (Trados Translator's workbench 1994)
- Database containing segments of translated content
- CAT tools are based on TM technology & concepts
 - Analysis (Weighted Word Count calculated from fuzzy match algorithm)
 - Segmentation
 - Translation
 - Term extraction
 - Concordance

Translation Memory



Advantages:

- Improves translation speed (long term projects)
- Savings for large projects
- Ensures better consistency/terminology
- Online collaboration (TM server/cloud)

Obstacles:

- Manual maintenance necessary to keep good quality
- Lower leverage for non-technical domains
- Could not generate long sentences well (the parts not always fit together)

Machine Translation



Advantages:

- Translate unseen content
- Consistent output
- Online translation

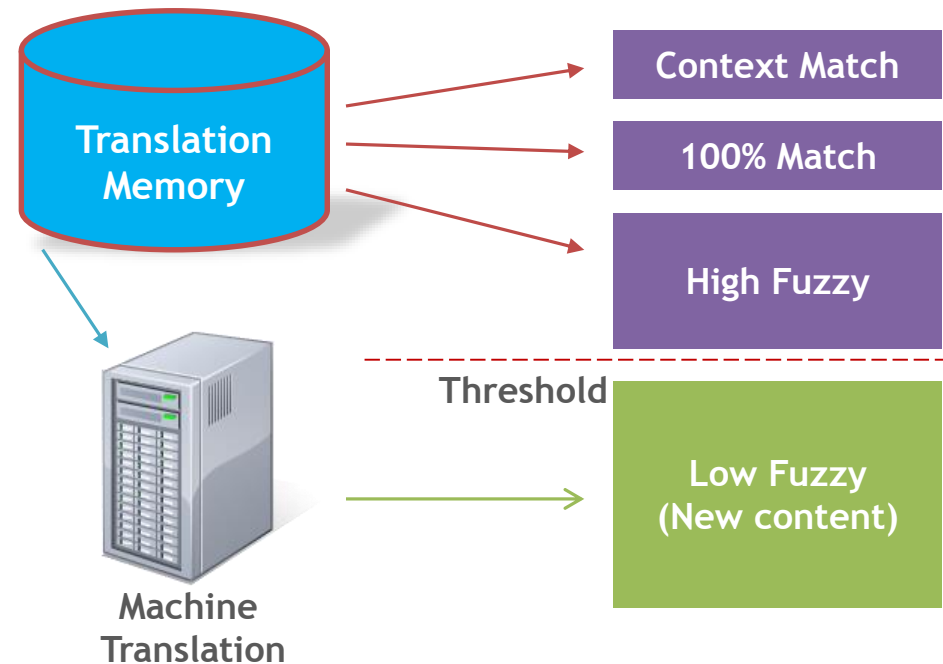
Obstacles:

- Quality differs among languages
- Need large corpus
- Non-transferable across domains

Combining Technologies : TM + MT

Goal to match source content against TM to plug in any previously-translated content, applying the MT/PE to only new strings.

Producing the Localized File



Why MT in the LSP industry at all?

Customers interested in:

- On-demand translations for daily content
- High volumes of content but limited budget to human translate everything
- Need for instant translation (e.g., chat)
- Right quality for right content type
- Short turnaround time (TAT) for translation

The background is a solid, vibrant red. Overlaid on this background are several thin, white, curved lines that sweep across the lower half of the image, creating a sense of motion and framing the text below. These lines are smooth and fluid, resembling a stylized signature or a decorative flourish.

Challenges of Applied MT

Challenges of Applied MT

- Integrating MT into the localization workflow
 - Connecting to CAT tools for efficient post-editor interaction with MT output
- Measuring the impact of MT on the localization process
 - Measuring Engine quality
 - Measuring post-editor productivity
- The impact of MT on the business relationship

Integrating MT into the localization workflow



- Internal General Purpose Tool
 - Operates on interchangeable loc. Formats (TMX, XLIFF)
 - MT technology Agnostic
- Connectors to CAT tools:
 - Native support (MS Hub, Google etc.)
 - MosesXMT - Moravia LetsMT compatible API (MemSource)
 - LTGear - global MT API (in development)

Importance of TMX

- File format for TM data interchange
- Bilingual
- Used in several scenarios throughout LSP process
 - CAT Tools
 - Integration with MT (for larger volumes)
 - Corpus creation
 - MT Training

Measuring the impact of MT (I)



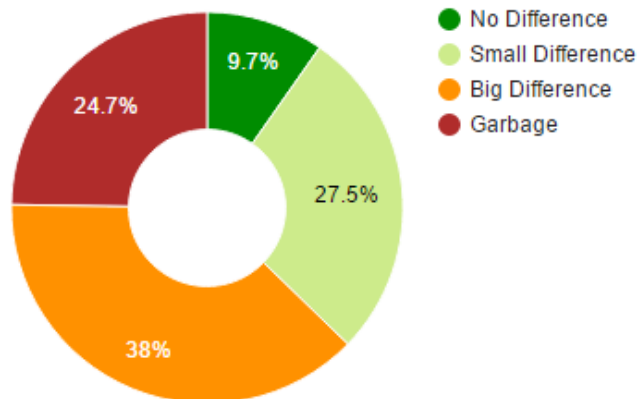
Moravia Platform for **EVA**luation of **MA**chine **T**ranslation

- Covers both Machine Contribution and Human Contribution to the final translation
 - Machine Contribution : Meteor
 - Human Contribution : RedBall (customized TER)
- Segment Level Testing A/B (SLAB)

Measuring the impact of MT (Ia) - Score Categorization

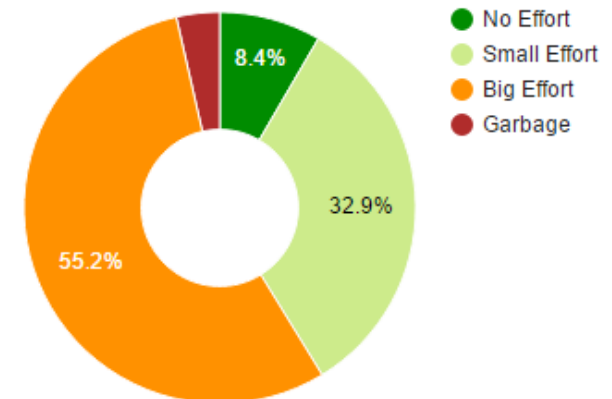
Meteor Score Categorization

Meteor Score Categorization Explained



RedBall TER Score Categorization

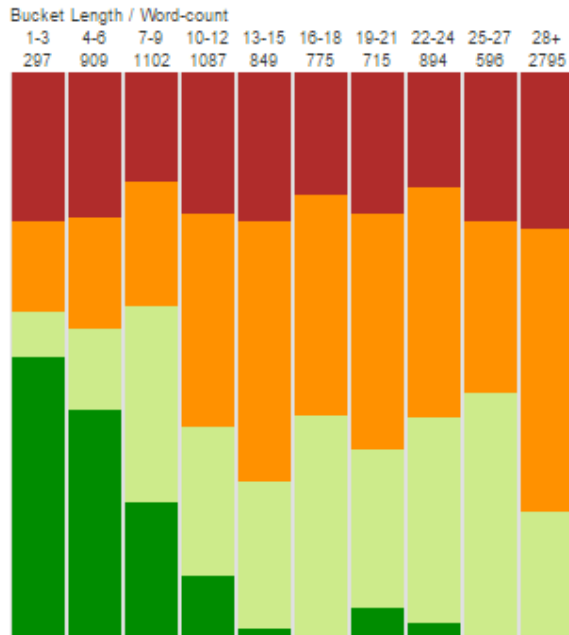
RedBall TER Score Categorization Explained



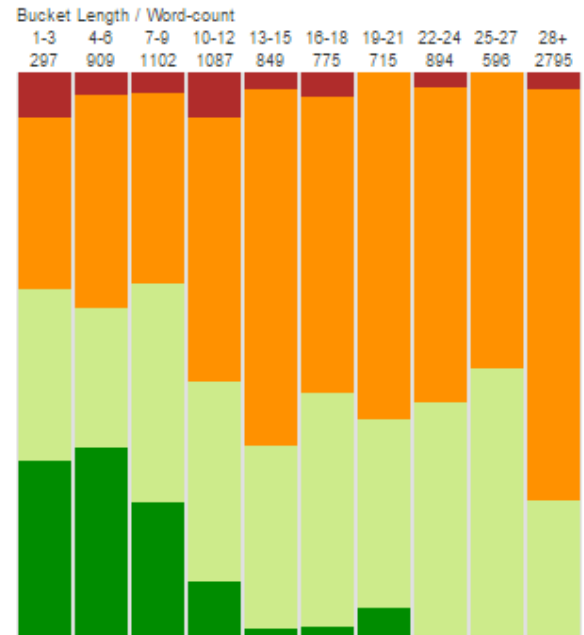
Measuring the impact of MT (Ib)

- Categorization By Segment Length

Segment Length Analysis (Meteor)



Segment Length Analysis (RedBall)

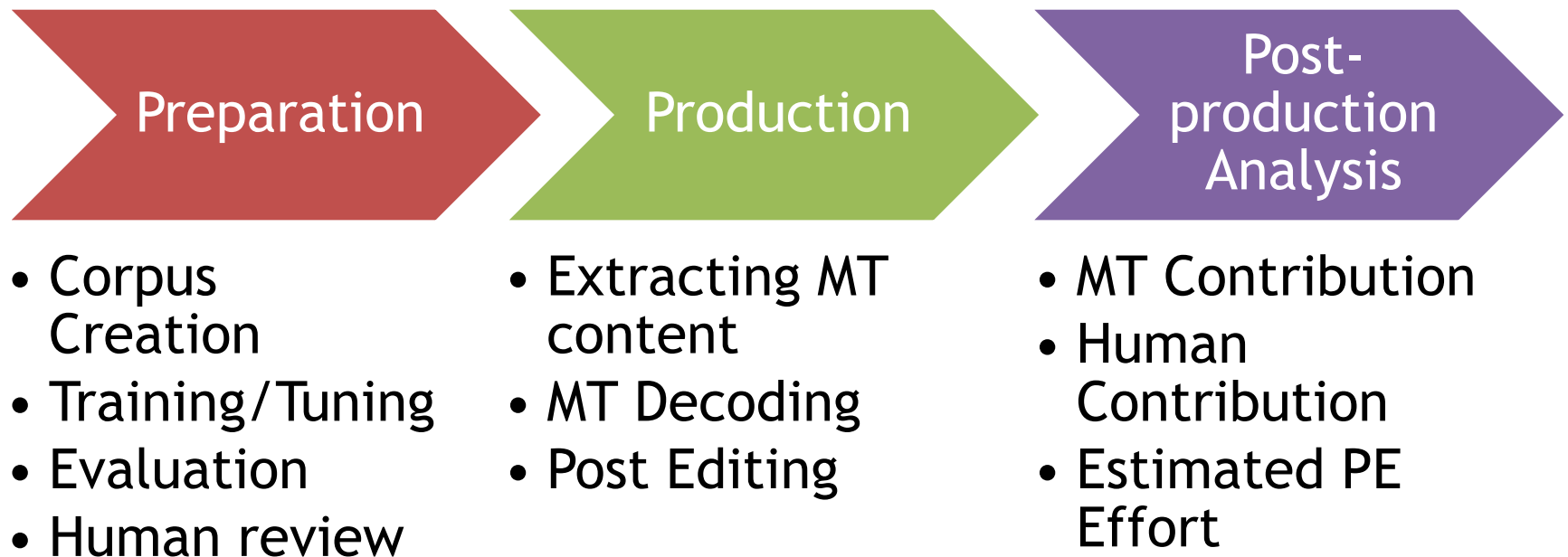


Measuring the impact of MT (II)

Human Evaluation

- Custom solution - Excel sheet processing
 - + Works offline
 - Manual distribution and collection
- Towards standardization:
 - TAUS's Dynamic Quality Framework(DQF)
 - WMT tools (Appraise)

Sample MT Engagement Process



MT Business negotiations

- Challenge - everybody thinks MT is free (e.g., Google)
- We approach MT as a productivity tool
- Important part of delivering a full suite of language services, not a standalone service or simple packaged solution
- Cost is not most important driver - it is speed / shorter turnaround time
 - Pricing is part of the business relationship. MT usage is only one of many driving factors

The image features a solid red background. In the lower-left quadrant, the text "Case Studies" is written in a bold, white, sans-serif font. The rest of the page is decorated with several thin, white, curved lines that sweep across the middle and right side, creating a sense of motion and flow. These lines are layered, with some overlapping others, and they generally curve from the left towards the right, ending in a more complex, overlapping pattern on the right side.

Case Studies

Case Study:

MT Program for Top Backup & Recovery Software Company

Challenge	Approach	Outcome
<ul style="list-style-type: none">• Continually cut down on localization costs• Improve turnaround time• Translate more content	<ul style="list-style-type: none">• Work with client to understand quality level mapped to content type• Develop workflows: MT+PE for tech docs, Raw MT for Knowledge Base• Select best of breed from multiple MT technologies<ul style="list-style-type: none">• MS Translator Hub wins• Evaluated outputs with automatic (EVAMAT) and human methods	<ul style="list-style-type: none">• All target content determined to be viable• Initial cost reduction 10%• Will increase to 20-30% reduction over time due to incremental retraining

Case Study:

Machine Translation Program for Automotive Diagnostic Tools

Challenge

- Stretch an existing localization budget
- Improve turnaround time
- Translate more content
- Low quality source content

Approach

- Understand quality level mapped to content type
- Tested Moses and MS Translator Hub; Moses prevailed
- Evaluated outputs with automatic (EVAMAT) and human methods
- Created terminology strategy: Moravia automated terminology tool to improve consistency

Outcome

- SW content didn't reach the required quality due to terminology issues
- Help + Doc content was viable with PE
- Initial time reduction minimal
- Source improvement improved MT output
- Corpus enhancements planned in future

Case Study:

Machine Translation and Post-Editing for Software Giant

Challenge

- Reduce global support costs and improve user experience
- Increase target languages, content types and volume
- Map content types to required levels of quality
- Get more value from the existing budget by improving productivity

Approach

- MS Translator Hub engine
- Optimize TMs to achieve higher leverage
- Full PE on high-profile content to achieve human quality
- Light PE on standard content to achieve acceptable “gisted” quality level

Outcome

- MT viable for 29 of 34 target languages, with average TM leverage of 72%
- Total program savings of 10%
- 1.5M machine-translated words produced in 2014
- Increased user self-support in more local markets; global support costs down by 10%+

Open topics

- MT Prediction
- More human-like Automated metrics
- Better MT for under-resourced languages
- Morphologically rich languages challenges
 - Czech, Korean, Finno-Ugric, Turkish
- Adding syntactic features into MT

Summary

- LSP usage of MT poses several challenges
- Presented solutions fitting in LSP environment
- Moravia has developed solutions for:
 - Integration
 - Impact Evaluation and Analysis
- Case studies representing different MT approaches
- Moravia is looking forward to participating in the MT Marathon

The image features a solid red background. In the lower half, there are three overlapping white ovals that create a sense of depth and movement. The word "Questions" is centered in white text within the space defined by these ovals.

Questions

Thank you!



Tomáš Fulajtár

tomasfu@moravia.com

+420 545 552 340