

Utilizing Language Data at an AI-driven Localization Technology Company

Aleš Tamchyna, Senior AI Research Manager



Language Data Space workshop, Prague, December 2 2024

Outline

Utilizing Language Data at an AI-driven Localization Technology Company

Introduction to Phrase

Brief overview of the company and its AI-based features

Utilizing data for AI

Phrase's approach to data acquisition and governance

Generative AI

New data-related trends and challenges in the GenAI era



Outline

Utilizing Language Data at an AI-driven Localization Technology Company

Introduction to Phrase

Brief overview of the company and its AI-based features

Utilizing data for AI

Phrase's approach to data acquisition and governance

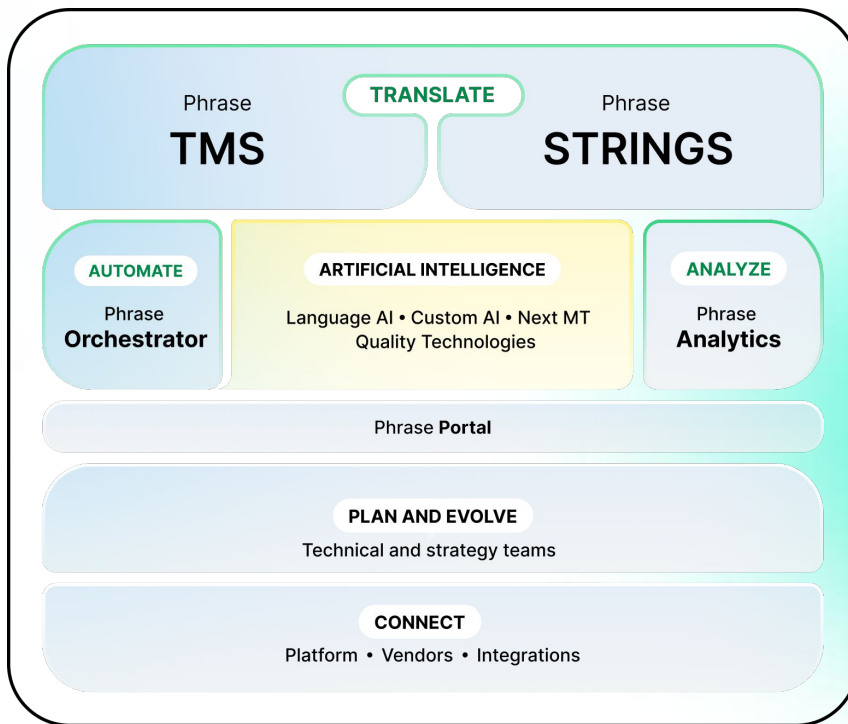
Generative AI

New data-related trends and challenges in the GenAI era



Phrase - Brief Introduction

- **Broad cloud-based Translation/Localization Platform**
- Focus on **Automation workflows**
- **Platform only** - for both Enterprises and LSPs
- **Strong Technology Aggregator** positioning
- **Customers are not siloed** to just one solution
- Allows dynamism in this time of **rapid transformation driven by GenAI**

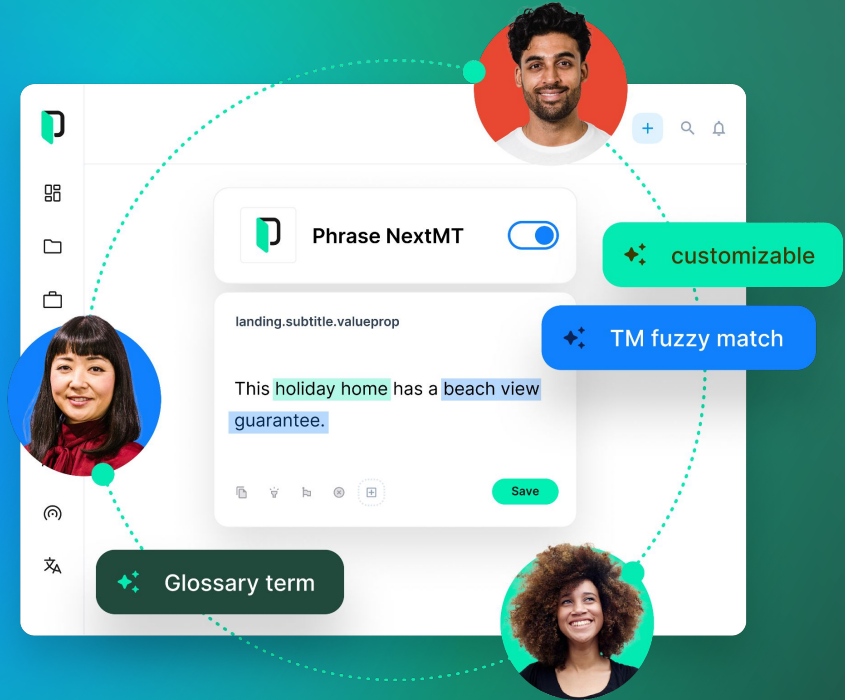


Phrase NextMT

In-house TMS ready MT engine

Key features:

- **Translation memory integration:** fuzzy matches from your translation memory improve MT translation accuracy by up to 50%.
- **Advanced glossary support:** forget about search and replace. NextMT ensures your terminology is always grammatically inflected.
- **Fully customizable:** with our Phrase Custom AI platform you can create your own model for unprecedented quality.

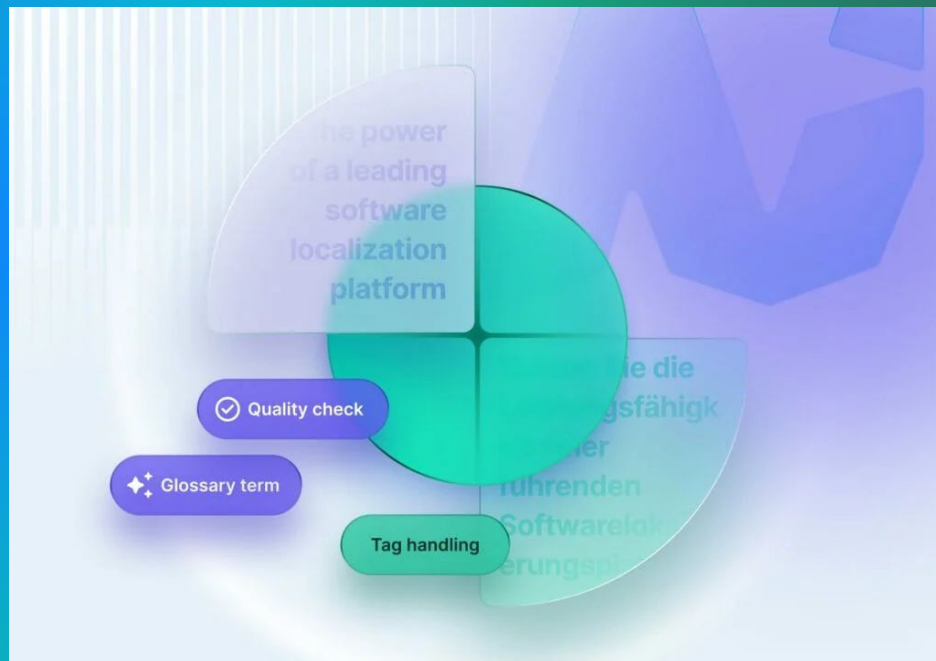


Phrase Next GenMT

NextMT Successor powered by OpenAI

Our new machine translation solution that fuses OpenAI's latest models with the design goals of Phrase NextMT.

Leverages MT glossaries, supports tag handling and *will* utilize few-shot prompting in our upcoming release.



Quality Performance Score

Quality Estimation for scalable monitoring and routing

QPS generates scores for machine translated segments from 0-100 using the MQM framework.

Scores enable:

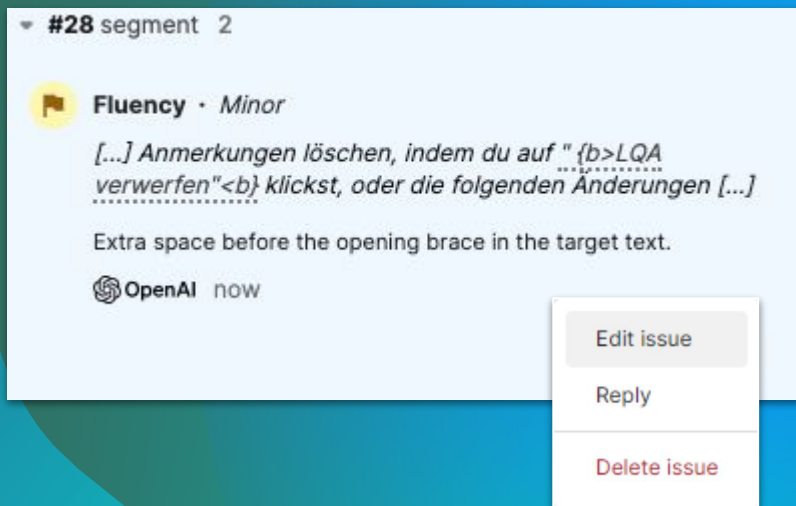
- **Post-editing**, providing linguists with segment-based quality scores
- **Project planning**, providing project managers with score overview for translation tasks
- **Custom workflows**, with user set quality thresholds

The screenshot displays a vertical toolbar on the left with icons for home, list, folder, clipboard, search, document, users, refresh, and a magnifying glass. The main content area shows two panels. The top panel is for the source language 'en' (English) and contains the text: 'Source: en', 'Check out our Help Center pages', and 'Watch our webinars'. The bottom panel is for the target language 'fr' (French) and contains: 'Target: fr', 'Consultez nos pages d'aide' with a yellow quality score of 75 and a warning icon, and 'Regardez nos webinaires' with a green quality score of 100 and a checkmark icon. A blue arrow points from the English text to the French text.

Language	Text	Quality Score	Status
en	Source: en		
en	Check out our Help Center pages		
en	Watch our webinars		
fr	Target: fr		
fr	Consultez nos pages d'aide	75	Warning
fr	Regardez nos webinaires	100	Success

Auto LQA

LLM-based Automated LQA



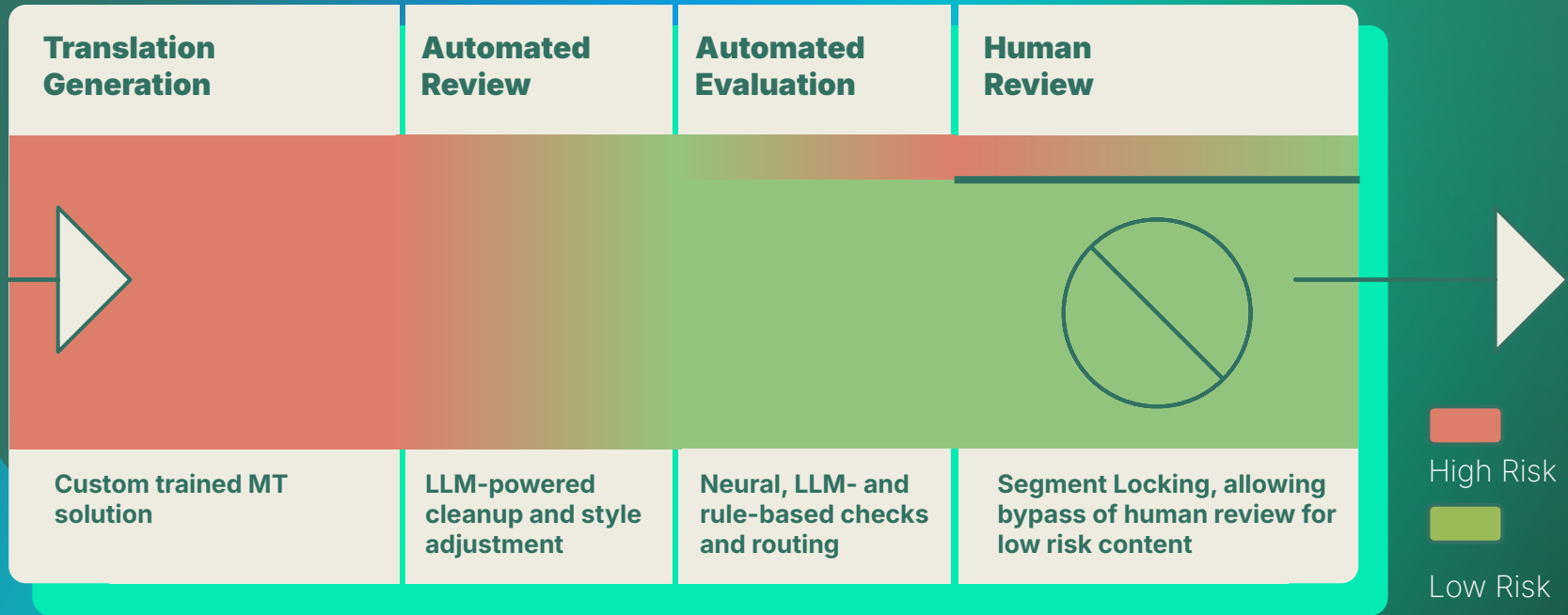
The screenshot shows a software interface for LQA. At the top, it says "#28 segment 2". Below that, there is a yellow flag icon followed by the text "Fluency · Minor". The main text of the issue is: "[...] Anmerkungen löschen, indem du auf " {b>LQA verwerfen"klickst, oder die folgenden Änderungen [...]". There are red dashed lines underlining the opening and closing braces of the code block. Below the text, it says "Extra space before the opening brace in the target text." At the bottom left of the issue card, there is the OpenAI logo and the word "now". A context menu is open over the issue card, containing three options: "Edit issue", "Reply", and "Delete issue".

Fully-automated alternative and complement to costly Human LQA, allowing **reduced costs and improved consistency**

- Perform a fully-automated MQM translation quality annotation:
- (Optionally) provide an explanation of the error and a suggested correction
- Standard MQM score is then calculated algorithmically based on the annotation

Auto Adapt

LLM-based Job-level Adaptation



Auto Adapt example

Source

Der wiederverwendbare Wasserflasche ist aus Edelstahl gefertigt und hält Getränke 24 Stunden lang kalt oder 12 Stunden heiß. Sie können die Flasche mit einem praktischen Griff einfach transportieren. Der Deckel ist **auslaufsicher und für die Spülmaschine** geeignet. Jeder Nutzer/in sollte sicherstellen, dass die Flasche regelmäßig gereinigt wird. Die Flasche ist in drei **Farben** erhältlich: Blau, Rot und Grün.

Wenn der Deckel beschädigt ist, **informieren Sie uns bitte, damit er so schnell wie möglich ersetzt werden kann.**

`auslaufsich(e|er|es|en) => leakproof`

Target (MT output)

The reusable water bottle is made of stainless steel and keeps drinks cold for 24 hours or hot for 12 hours. You can easily carry her with her handy handle. The lid is **spill-proof and dishwasher safe**. The bottle is available in three **colors**: blue, red, and green. Each user should make sure he cleans it regularly.

If the lid is damaged, please inform us so that it can be replaced as soon as possible.

1. Inconsistent between segments
2. Includes linguistic errors
3. Doesn't follow terminology resource
4. Mixed formality

Target (Auto Adapt output)

The reusable water bottle is made of stainless steel and keeps drinks cold for 24 hours or hot for 12 hours. You can carry it easily with its convenient handle. The lid is **leakproof and dishwasher-safe**. The bottle is available in three **colours**: blue, red, and green. Users should make sure they clean it regularly.

If the lid is damaged, let us know and we'll replace it straight away.

1. Improve document consistency
2. Improve fluency
3. Fix terminology errors
4. Specify formality level
5. Add custom instructions

Outline

Utilizing Language Data at an AI-driven Localization Technology Company

Introduction to Phrase

Brief overview of the company and its AI-based features

Utilizing data for AI

Phrase's approach to data acquisition and governance

Generative AI

New data-related trends and challenges in the GenAI era



Powering AI Features with Data

Balancing data privacy and functionality

“Non text generating” models

- Quality estimation
- Detection of non-translatable segments
- MT Autoselect

Both public and aggregate customer data

- No potential for data leaks, no privacy issues

Text generating models

- Machine translation

Nuanced approach

- Generic models
 - Public or purchased data *only*
- Customized models
 - Utilize data of the specific customer
 - Customer is in control



Common challenges

- Data retention requirements
- Anonymization
 - Desirable in principle, but can't be implemented naively
- Regulatory and legal constraints



Data marketplaces - impact on Phrase

As a **client**:

- Phrase purchased significant amounts of language data for various purposes
 - Improving MT quality with additional training data
 - Expanded language/domain coverage for MT evaluation
- Mixed experience
 - Customer data so far more useful in providing value to our customers

As a **provider**:

- Limited potential
 - Vast majority of data at Phrase is proprietary and belongs to our customers
- We have released annotated datasets based on publicly available data



Outline

Utilizing Language Data at an AI-driven Localization Technology Company

Introduction to Phrase

Brief overview of the company and its AI-based features

Utilizing data for AI

Phrase's approach to data acquisition and governance

Generative AI

New data-related trends and challenges in the GenAI era



Impact on customer perception of AI

- Major increase in awareness
- Core challenge: **communicating the value of providing data to our customers**
- Promising strategy: prefer real-time adaptation to training/fine-tuning
 - Avoids “capturing” customer data in trained models



Evolving requirements

- Major areas:

- Robustness
- Responsible AI
- AI risk management
- Governance, ML ops



Data in the Generative AI era

- Decreasing value of large-scale data sources
- Data sources/corpora with **high added value** becoming crucial
 - Expert linguistic quality annotations
 - Very high-quality texts and translations
- Increased importance of expert-annotated sets for **evaluation**
 - Fitness for purpose > general quality



Takeaways

Utilizing Language Data at an AI-driven Localization Technology Company

- Technology companies such as Phrase increasingly rely on data to realize value for customers.
- Increased awareness about data and AI among customers and regulators alike.
- Generative AI dramatically transforms data needs and requirements.



Thank you!

Questions?