産業翻訳における機械翻訳の継続的改善手法

株式会社ヒューマンサイエンス　中山雄貴

HUMAN SCIENCE

MTrans

ポストエディターの作業

1. 翻訳メモリから流用

原文を読み、過去の訳文を確認・修正する

2. ポストエディット

原文を読み、機械翻訳の訳文を確認・修正する

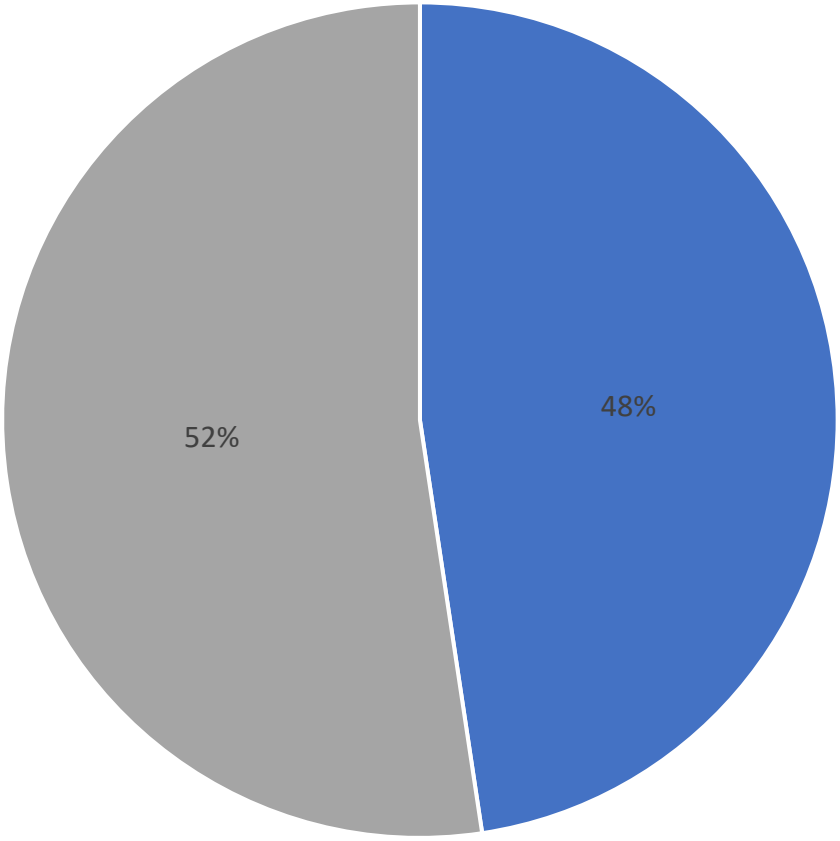
ポストエディターの作業

1. 翻訳メモリから流用

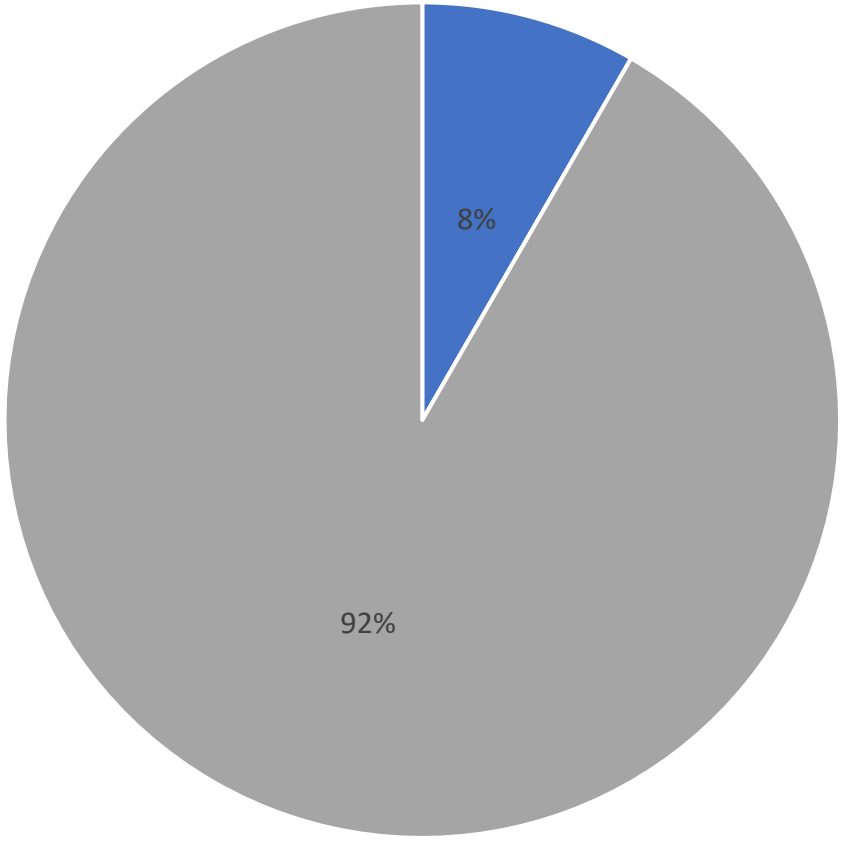
原文を読み、過去の訳文を確認・修正する

流用文の割合：新規文書の場合

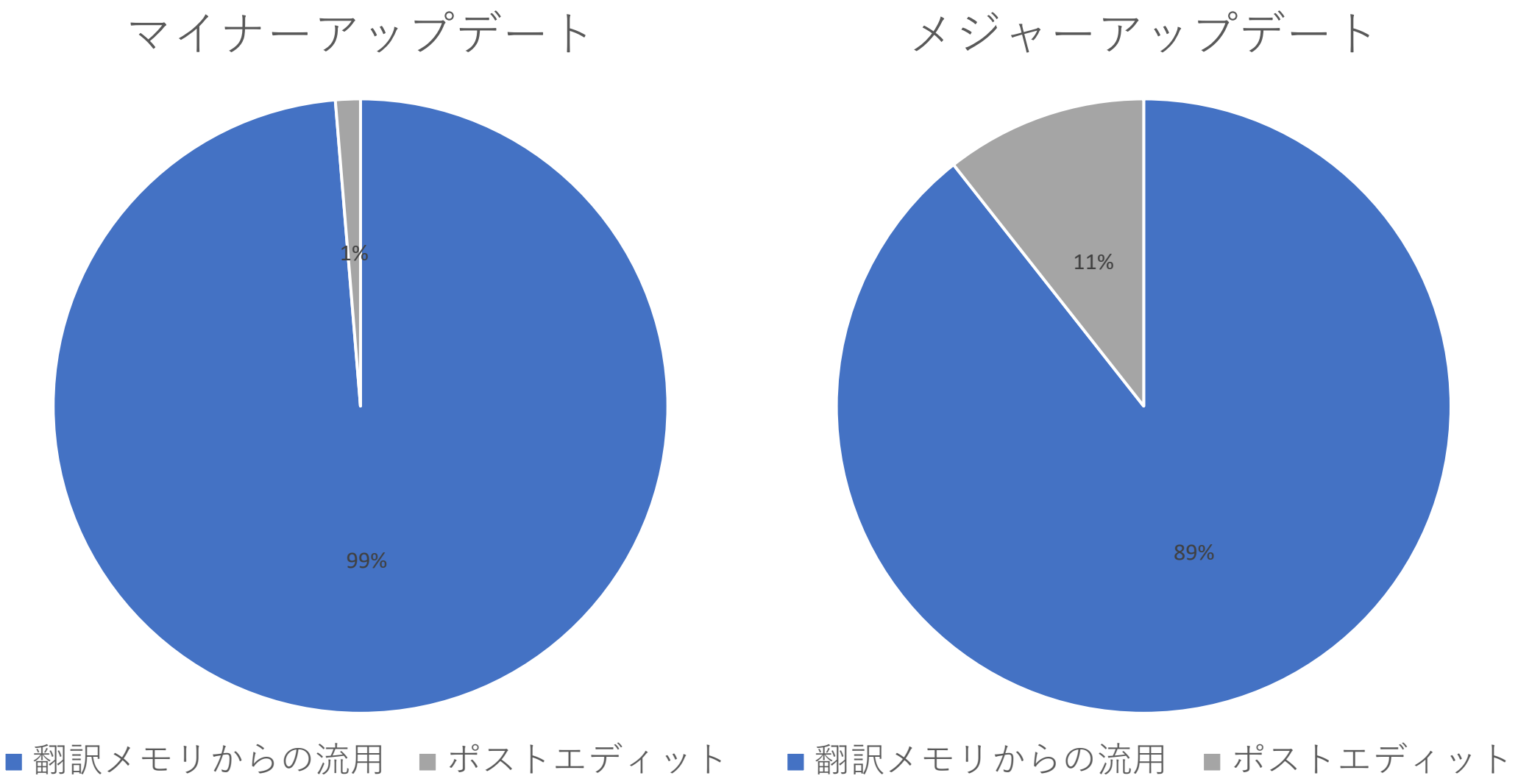
定型表現あり



定型表現なし



流用文の割合:改訂文書の場合



翻訳メモリの種類

翻訳用

- 一時的：プロジェクトごとに作成、破棄される
- 翻訳者が読み書きする

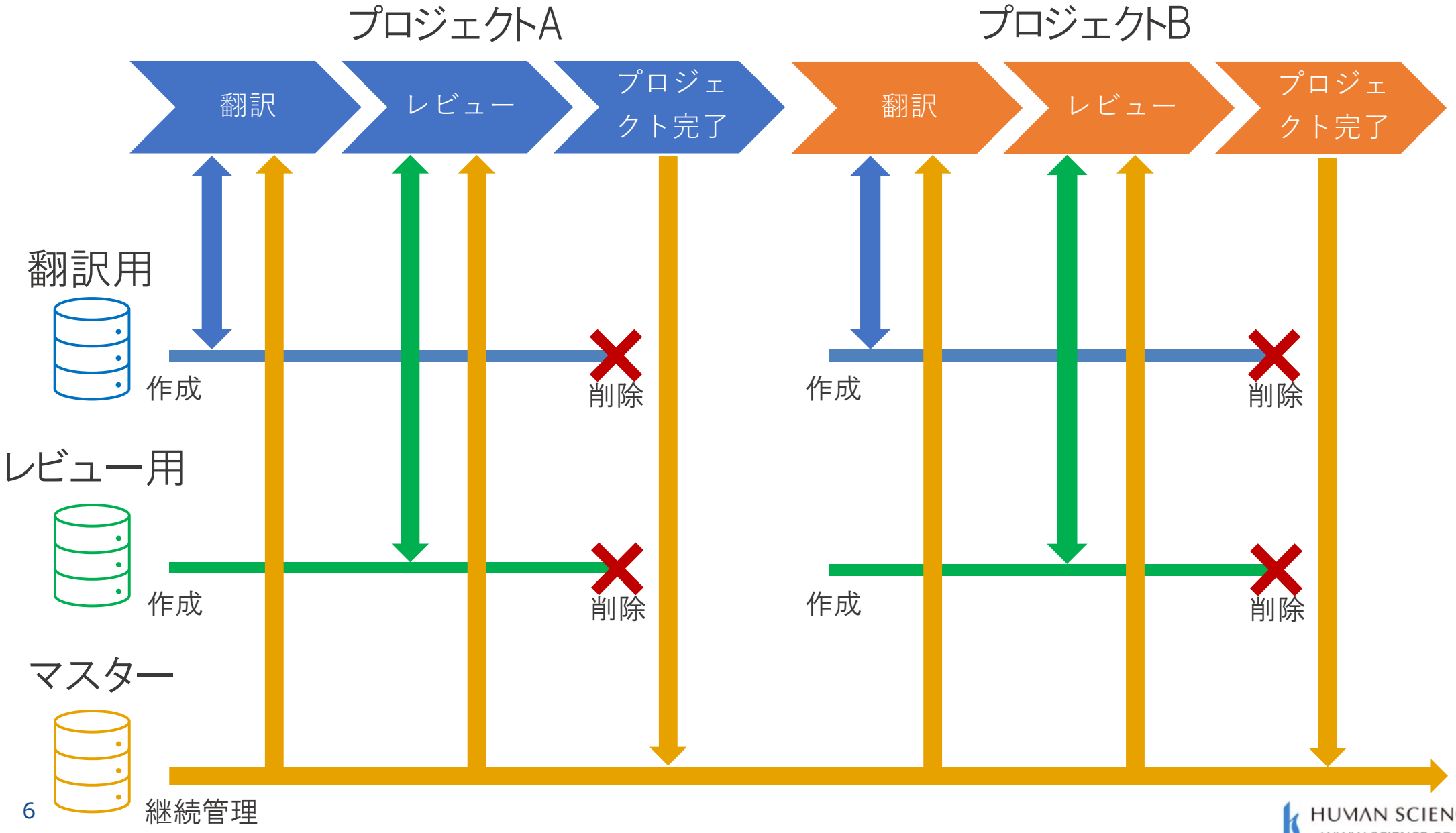
レビュー用

- 一時的：プロジェクトごとに作成、破棄される
- レビューアが読み書きする

マスター

- 永続的：プロジェクトを超えて使用される
- 翻訳者とレビューアが参照専用で使用する
- プロジェクトマネージャがプロジェクト完了時に書き込む

翻訳メモリのライフサイクルと読み書き

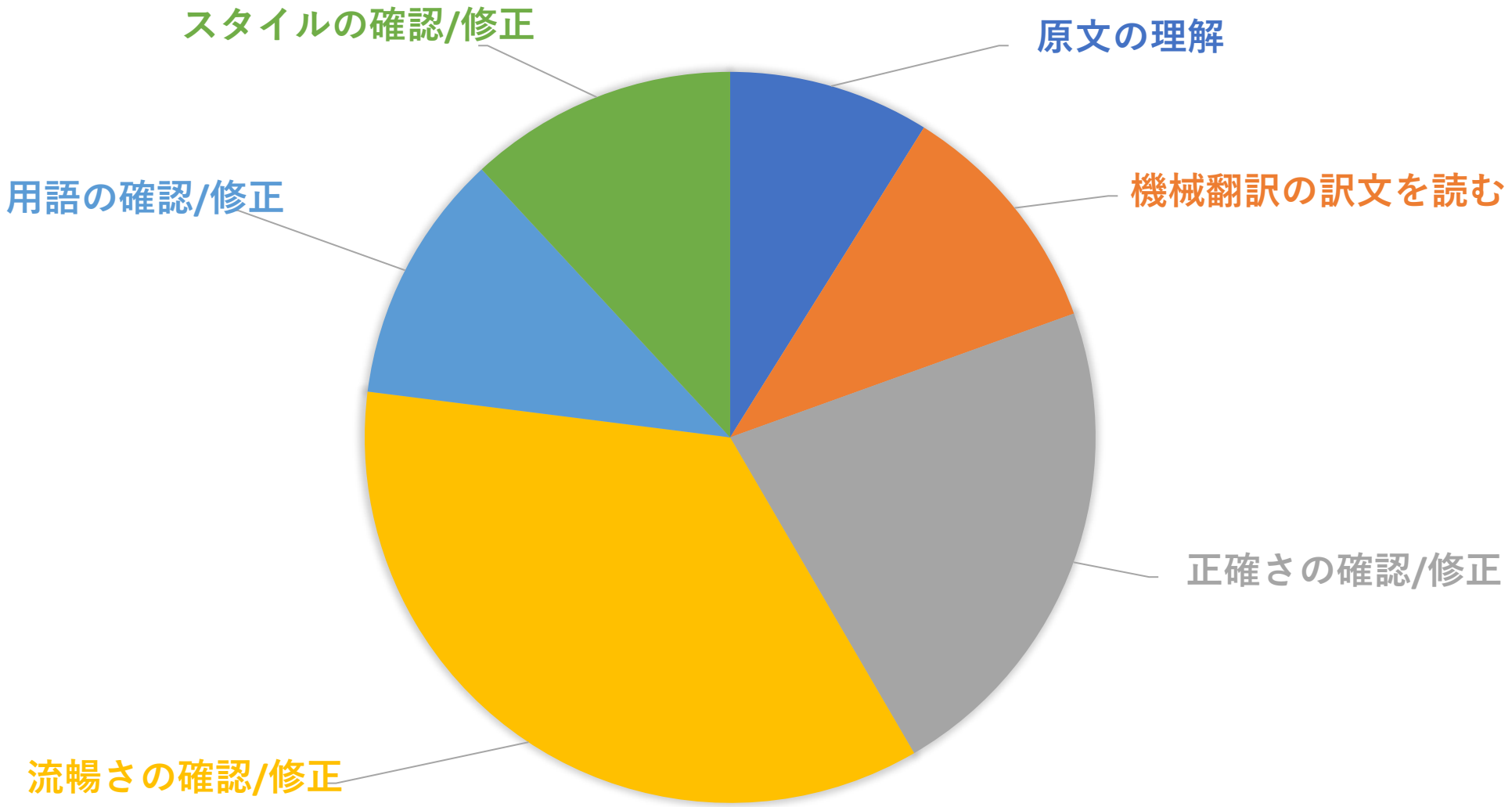


ポストエディターの作業

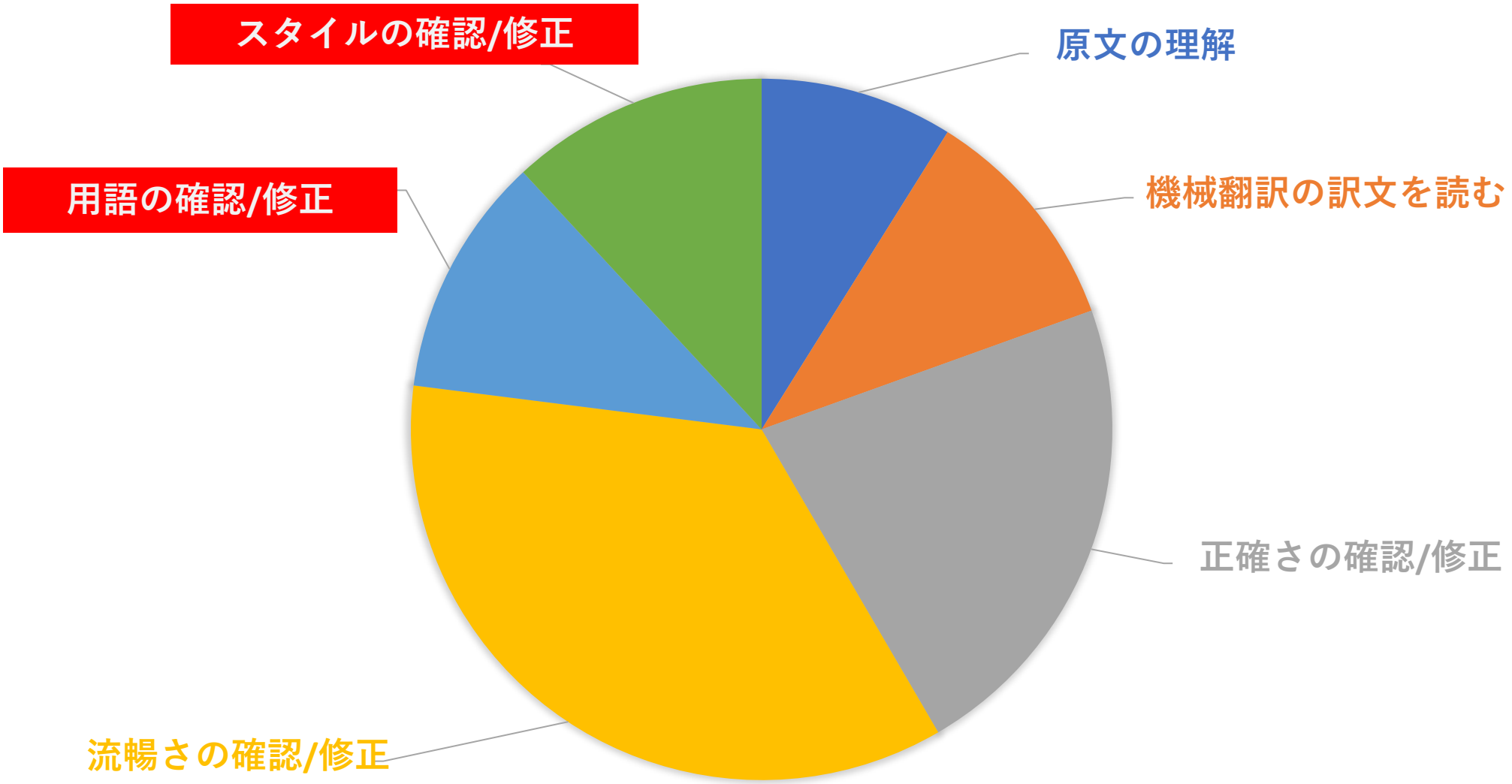
2. ポストエディット

原文を読み、機械翻訳の訳文を確認・修正する

ポストエディットの作業内容の割合



ポストエディットの作業内容の割合



用語とスタイルの誤り

機械翻訳エンジンが出力する素の訳文には用語やスタイルの誤りが多く含まれ、単純な修正作業が多い

The file is saved in C:\Users\username\Downloads.	AT	ファイルはC：¥ Users ¥ username ¥ Downloadsに保存されます。
HUMAN SCIENCE provides machine translation solutions.	AT	ヒューマンサイエンスは、機械翻訳ソリューションを提供しています。
Please contact our Business Development department for details.	AT	詳細については、事業開発部門にお問い合わせください。
PE booster is a software to support post editors.	AT	PEブースターは、投稿編集者をサポートするソフトウェアです。

スタイル … コロンが全角。不要なスペースがある。
用語 … 用語集とは異なる訳語が使われている。

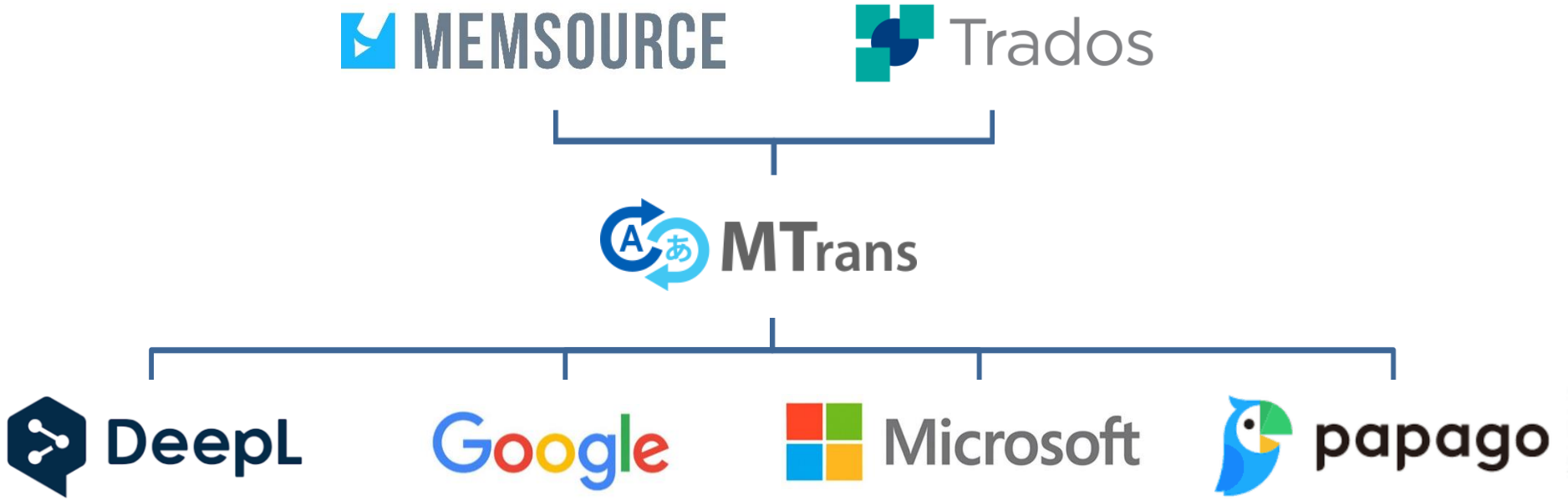
The file is saved in C:\Users\username\Downloads.	AT	ファイルはC:¥Users¥username¥Downloadsに保存されます。
HUMAN SCIENCE provides machine translation solutions.	AT	HUMAN SCIENCEは、機械翻訳ソリューションを提供しています。
Please contact our Business Development department for details.	AT	詳細については、事業推進部にお問い合わせください。
PE booster is a software to support post editors.	AT	PE Boosterは、ポストエディターをサポートするソフトウェアです。

スタイル … コロンを半角に修正。不要なスペースを削除。
用語 … 用語集の訳語へ修正。

機械翻訳サービスの用語集・スタイル置換機能

用語とスタイルの修正作業を自動化

DeepLなどの機械翻訳エンジンに汎用的に追加できる用語集・スタイル置換機能を開発し、
Memsource、Tradosから利用できるようにしました。

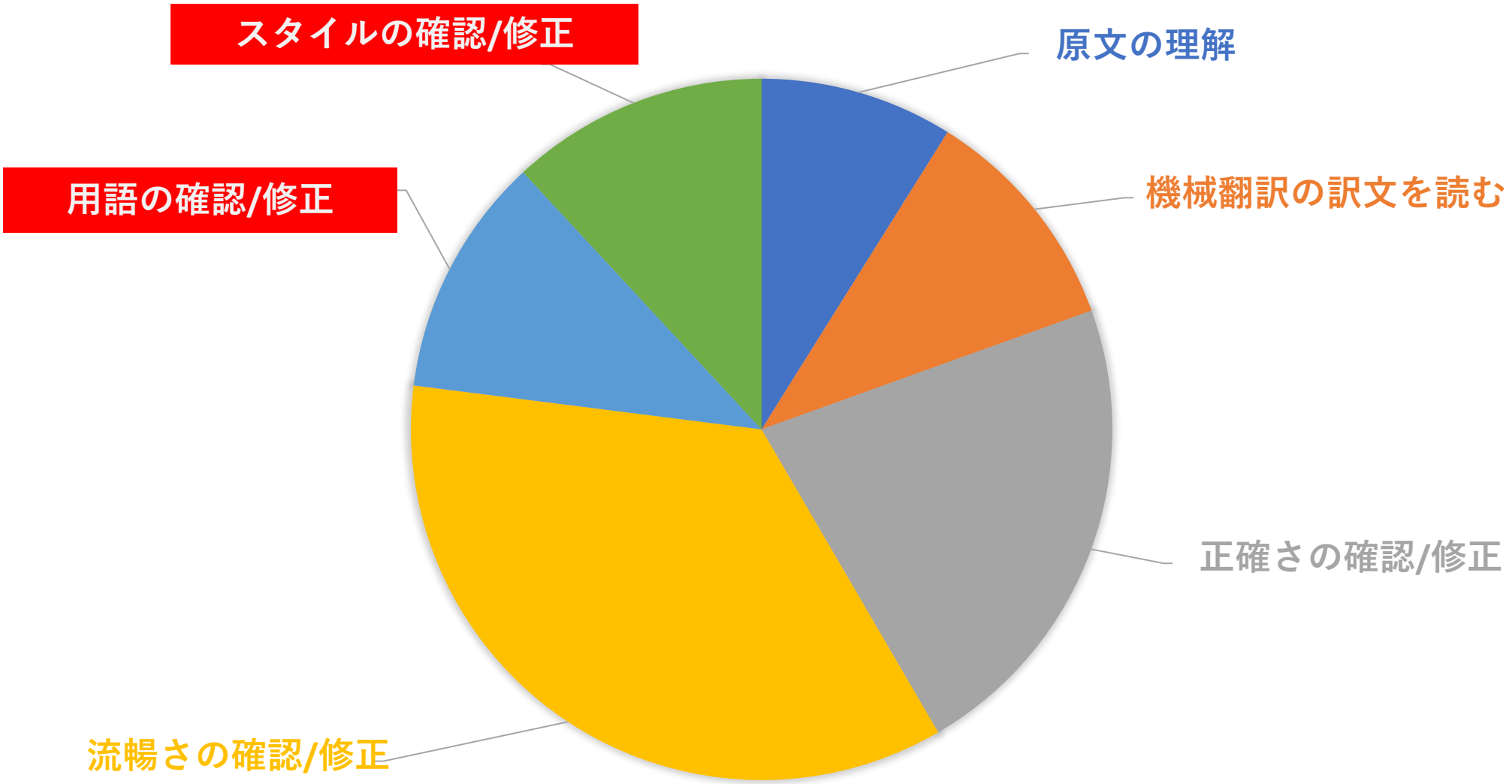


自動化によるポストエディット時間の削減

用語とスタイルの修正を自動化することで、ポストエディット時間が削減されます。

Google		Google + MTrans
109文中95文 87.2%	誤りが存在した 文の数	0文 0%
661	誤りの数	0
25分5秒	修正にかかった 時間	0秒
単純作業が多く手間に感じる ☹	ポストエディターの 感想	翻訳に集中できる ☺

ポストエディット時間を20%以上削減



用語集の注意点

ポストエディターが使用する用語集と、機械翻訳が使用する用語集は別々に管理します。

機械は文脈を理解できません。用語集の中にある訳語を、文脈に応じて使う・使わないを判断したり、複数の訳語の中から適切なものを選択したりすることはできません。

機械翻訳用の用語集には、無条件で使用する用語のみを登録します。用語ごとに訳語を1つのみ登録します。

また、用語を指定できるのは固有名詞と一般名詞のみです。動詞や文を登録することはできません。

英語	日本語
assertion	アサーション（開発関連）
battery	電池、バッテリー

英語	日本語
~~assertion~~	~~アサーション~~
~~battery~~	~~電池、バッテリ~~

用語集とスタイル置換条件の継続的改善

プロジェクト終了時に、ポストエディット前後の差分を作成し、保存しておきます。
定期的にこの差分を確認し、用語集とスタイル置換条件を更新します。

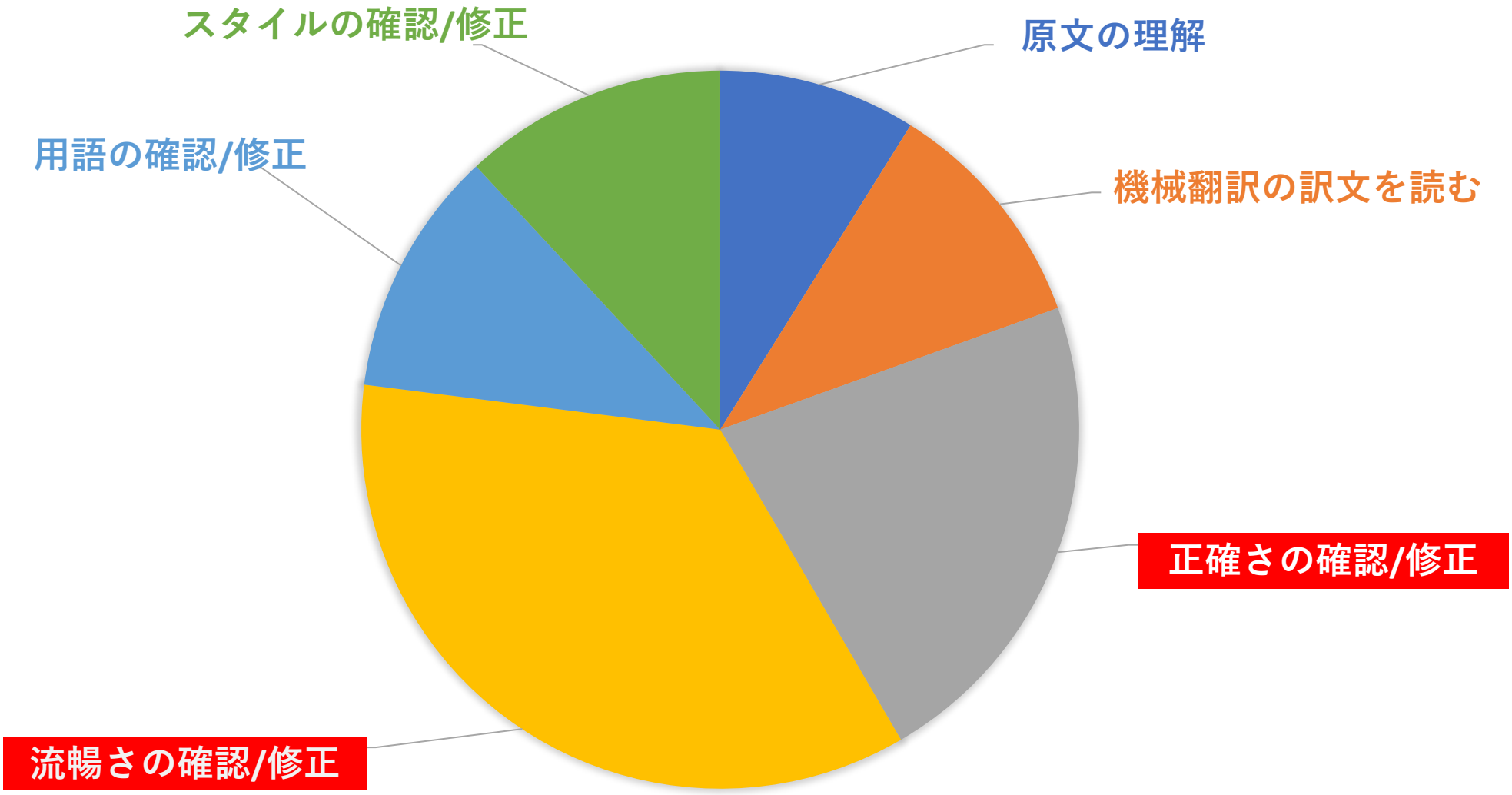
MEMSOURCE

標準機能に差分作成機能あり。
ヘルプの「Evaluating MTQE Results」を参照。

Trados

無料のプラグイン「Post-Edit Compare」を利用。
RWS AppStoreからダウンロード可能。

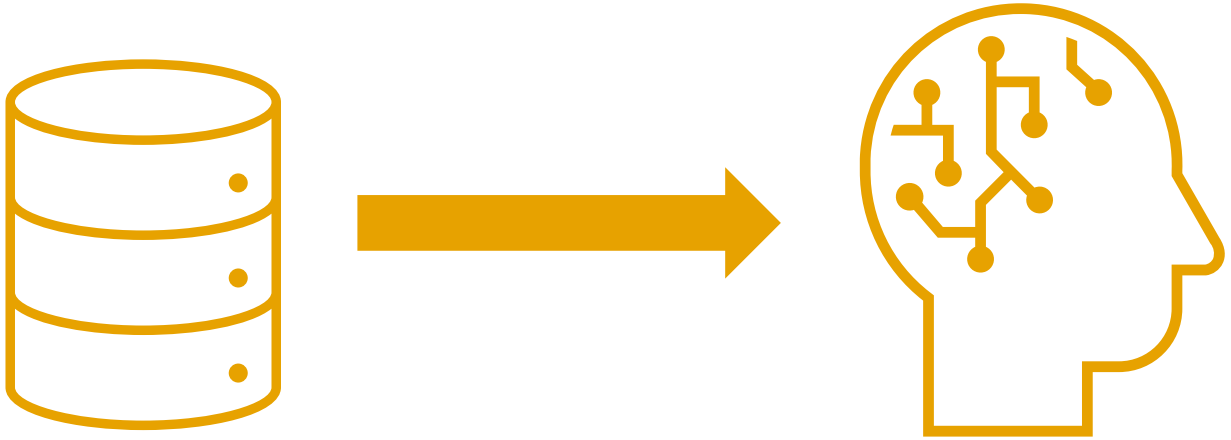
ポストエディットの作業内容の割合



機械翻訳モデル学習

モデル学習とは

汎用の機械翻訳モデルに対して独自の原文・訳文のペアを学習(トレーニング)させ、
独自の訳文を生成できるようにすること。



モデル学習のメリット

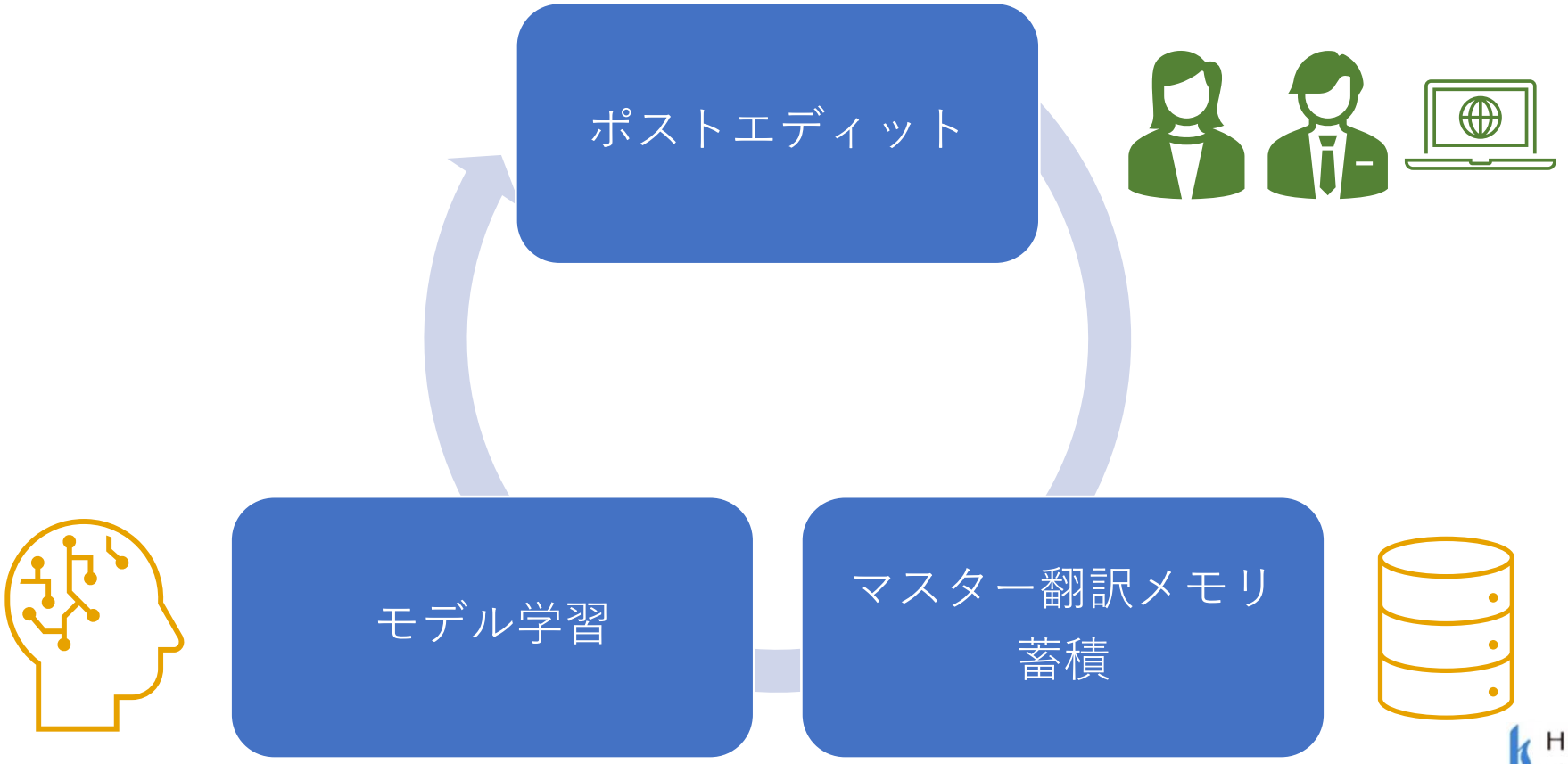
学習した分野の訳文の精度が向上します。
Google AutoMLの検証ではBLEU値が3.1ポイント向上しました。

モデル学習前	39.7
モデル学習後	42.8 ⬆3.1

分野:IT
言語:英日
学習データ分量:800文

モデル学習に使用するデータ

モデル学習には、マスター翻訳メモリに保存されている原文と訳文のデータを使用します。日々の翻訳案件で質の良い原文・訳文のペアを蓄積することが、高品質なモデル作成に繋がります。



モデル学習の注意点

学習データの訳文、用語、スタイルがそのまま出力されるようになるわけではありません。

翻訳メモリと用語集・スタイル置換機能を引き続き利用する必要があります。

原文	The NIC exists on the ‘Data Link Layer’ (Layer 2).
学習データの訳文	NIC▲は「データリンク層」▲(第▲2▲層)▲に位置します。 （「▲」は半角スペースを示しています）
モデル学習前の訳文	NICは、「データリンク層」（層2）に存在します。
モデル学習後の訳文	NICは、「データリンク層」（レイヤー2）に存在します。

まとめ

マスター翻訳メモリの
維持管理

ポストエディット前後の
差分を使った
用語集・スタイル置換条件の更新

マスター翻訳メモリを
使ったモデル学習

お問い合わせ

機械翻訳の導入・活用をお手伝いします。

ホームページからお問い合わせください。

www.science.co.jp/nmt

機械翻訳に関するブログ記事も掲載しています。

DeepL翻訳で機密は保持される？セキュリティは？

医療翻訳でのDeepLの翻訳精度は？Google/Microsoft/Amazonとの比較結果

韓国語が得意なAI翻訳「NAVER Papago」