WITAJ Rěcny centrum
Rěčny centrum
Sprachzentrum
Language Centre

Online-Übersetzer Deutsch-Sorbisch und Sorbisch-Deutsch (ODS) –

Herausforderungen und Lösungen

WITAJ Rěcny centrum Rěčny centrum Sprachzentrum Language Centre

Inhalt

1. Entstehung des Projekts mit dem Microsoft Translator Hub
2. Aktuelle Lösungen mit MOSES-Decoder und OpenNMT
3. Ausbau des zweisprachigen Korpus

1. Entstehung des Projekts mit dem Microsoft Translator Hub

zwei sorbische IT-Experten und Enthusiasten Wito Böhmak und Bernhard Baier

→ Kontakt zu Microsoft – Kontaktperson Helge Schroda

→ sorbische Buchstaben und Rechtschreibkontrolle in Microsoft Office + soblex

2015 gab Helge Schroda den Anstoß, einen zweisprachigen Korpus aufzubauen, und richtete im Microsoft Translator Hub ein sorbisches Übersetzungsprojekt ein

→ im WITAJ-Sprachzentrum werden seit 2016 zweisprachige Texte gesammelt und bearbeitet

1. Entstehung des Projekts mit dem Microsoft Translator Hub

2016 lernte Beata Brězanowa den IT-Experten Olaf Langner beim Sorbischen Stammtisch in Bautzen kennen

→ er erforschte die Funktionalität des Translator Hubs professionell

→ im WITAJ-Sprachzentrum wurden weiter zweisprachige Texte gesammelt und verfasst, bis eine erste Übersetzung möglich war

→ die Qualität der Übersetzungen hat das WITAJ-Sprachzentrum ermutigt, einen Projektantrag einzureichen – Beginn 01.01.2019

→ Marion Kremer aus Seattle wurde Beraterin bei linguistischen Fragen

1. Entstehung des Projekts mit dem Microsoft Translator Hub

→ im April 2019 sperrte Microsoft den Zugang zum Translator Hub

→ Enttäuschung und neue Herausforderung, den bereits erreichten Stand zu ersetzen

Internat ma 57 **městnow** za šulerki a šulerjow wot 5. do 12. lětnika.
Wjetšina z nich **chodźi** do 9. abo wyšeho lětnika.

Wšitcy šulerjo maja domjacy **porjad** dodźeržeć, při čimž płaća za młódše **holcy a hólcy** hinaše časy wuchoda a nócneho **měra** hač za staršich.

Móžnosć za šulerjow, internatne žiwjenje sobu postajeć a **wobchad z prawidłami wobwliwować**, skići internatna rada.

Das Internat hat 57 **Arbeitsplätzen** für Schülerinnen und Schüler der 5. bis 12. Klasse.

Die Mehrzahl von Ihnen **wandelt** in 9. oder des Klasse.

Alle Schüler verfügen Hausgenossen **Reihenfolge** eingehalten werden, wobei gelten für die jüngeren **Mädchen und jungen** andere Zeiten **wuchoda** und **nócneho des Friedens** als für die Eltern.

Die Möglichkeit für Schüler, **internatne** Leben mit **postajeć** und **Umgang mit Regeln zu beeinflussen**, angeboten wird **internatna** Rat.

1. Entstehung des Projekts mit dem Microsoft Translator Hub

From: Marion Kremer
Sent: Sunday, November 3, 2019 5:39 PM

Liebes Team,

das Team bei Microsoft hat mir jetzt bestätigt, dass es den Bau des Sprachmodells für NMT auf Q4 (April-Juni 2020) verschoben hat, wir sind also noch im Spiel. Wenn die Zeit näher rückt, schauen wir, auf wie viele Satzpaare Ihre Sammlung auch mithilfe der zusätzlichen Ressourcen angewachsen ist und wie wir strategisch weiter verfahren.

Herzliche Grüße gerade noch aus Seattle,

Marion Kremer

1. Entstehung des Projekts mit dem Microsoft Translator Hub

- es ist sinnvoll, den Kontakt zum Microsoft-Team weiter zu pflegen, zumindest solange die Chance besteht, dass Sorbisch in den Bing Microsoft Translator integriert wird

- derzeit steht aber die Recherche nach Alternativen im Mittelpunkt und darüber berichtet nun der IT-Experte Olaf Langner

2. Aktuelle Lösungen mit dem MOSES-Decoder und OpenNMT

→ <u>Grundlage:</u> großer bilingualer Korpus mit sorbischen und deutschen Sätzen

→ Sätze werden unterschiedlich analysiert, geteilt und wieder zusammengefügt

→ <u>aktuelles Tool:</u>

statistisches System

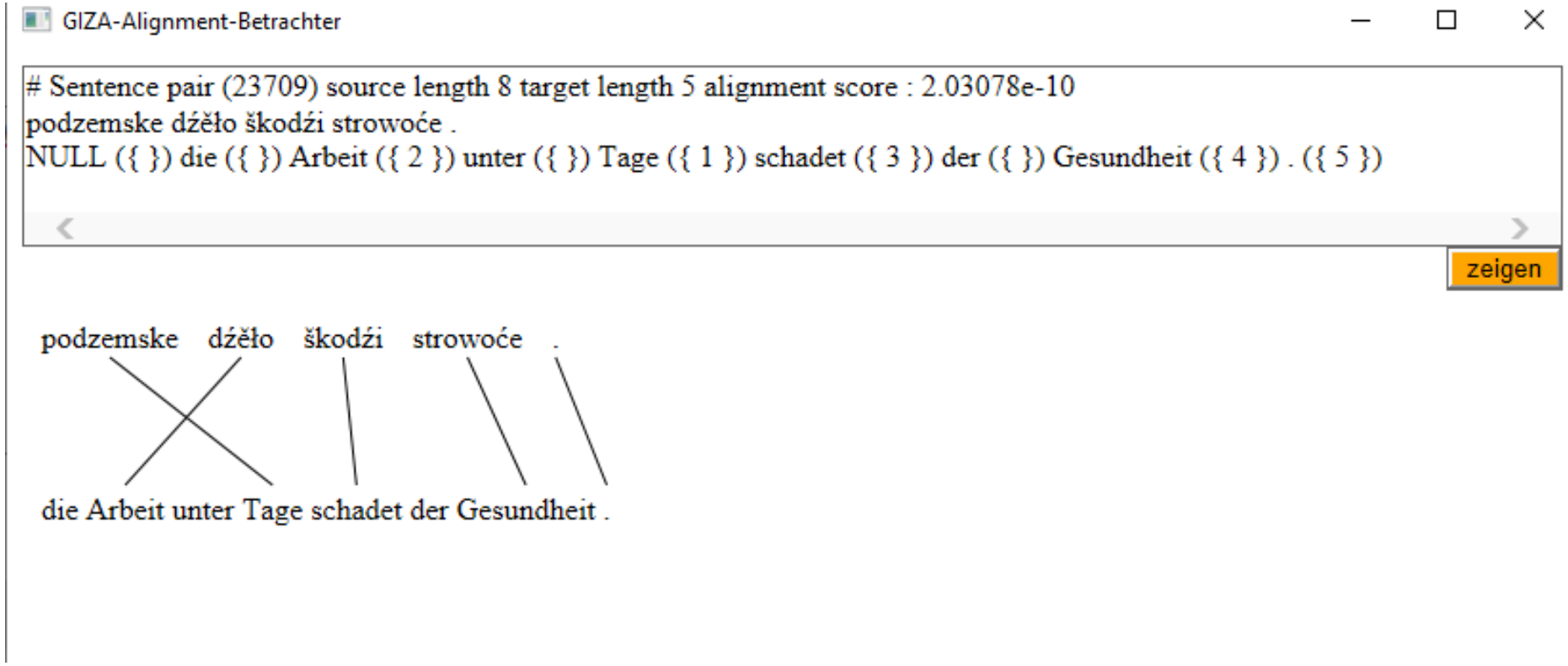
MOSES-Decoder

→ <u>paralleles Tool:</u>

neuronales System

OpenNMT (Python)

2. Aktuelle Lösungen mit dem MOSES-Decoder und OpenNMT



→ MOSES-Decoder

2. Aktuelle Lösungen mit dem MOSES-Decoder und OpenNMT

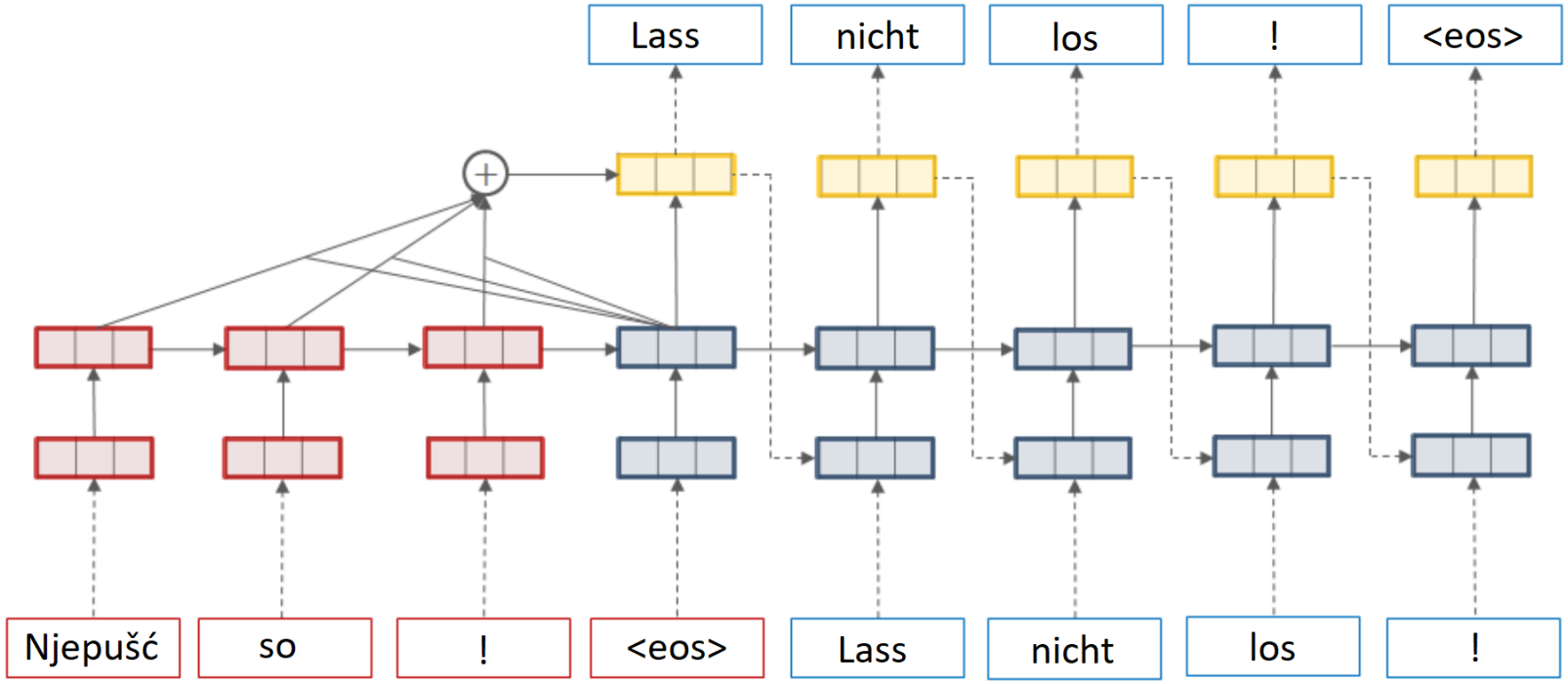


Figure: Sequence-to-sequence model with attention.

→ OpenNMT

2. Aktuelle Lösungen mit dem MOSES-Decoder und OpenNMT

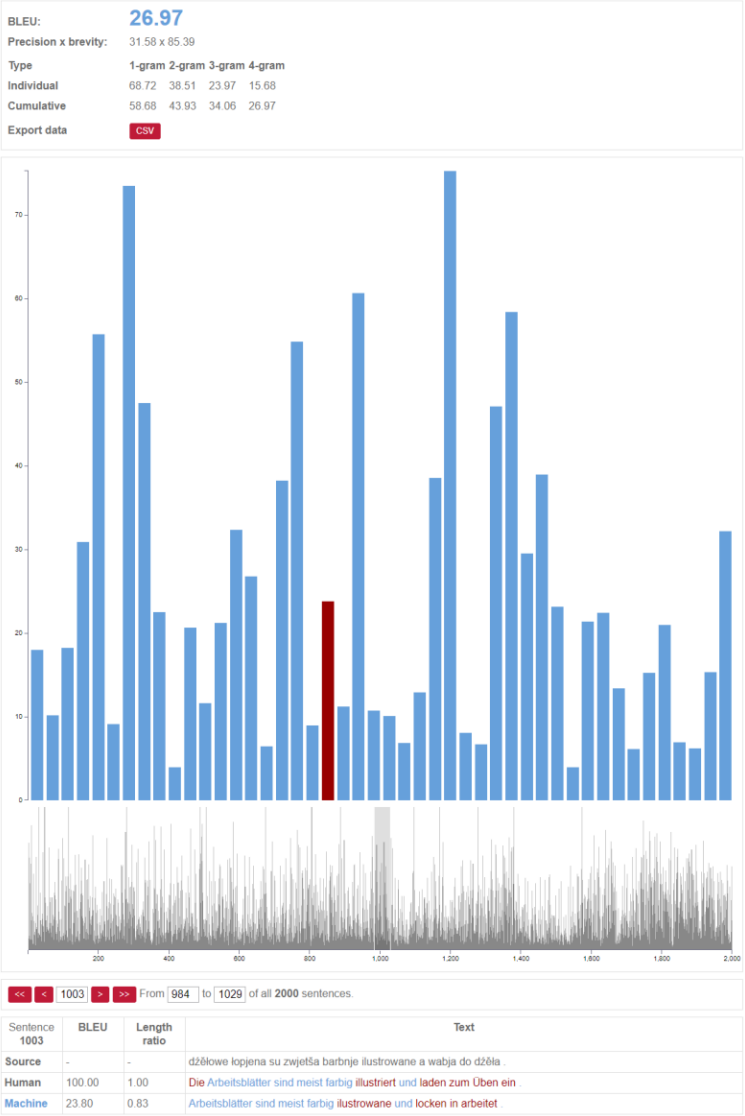
→ ein mit dem MOSES-Decoder übersetzter Satz im Vergleich zur Originalübersetzung

Quelle	→	Hdyž so hodźina započa, dyrbjachmy stańši prawu packu zběhnyć a z wotpowědnym postrowom strowić.
mensch. Übers.	→	Zu Beginn der Stunde mussten wir stehend die rechte Hand erheben und mit entsprechendem Gruß grüßen.
masch. Übers. mit dem MOSES-Decoder	→	Wenn Stunde begann, mussten stańši rechte packu die und mit entsprechendem Gruß grüßen.

2. Aktuelle Lösungen mit dem MOSES-Decoder und OpenNMT

→ BLEU (bilingual evaluation understudy)

→ Algorithmus zur Bewertung von Übersetzungen



3. Ausbau des zweisprachigen Korpus

→ Grundlage: zweisprachige Sätze, die täglich produziert werden, öffentlich zugänglich sind und mit vorgegebenen Begriffen verfasst werden

→ die Sätze werden so bearbeitet, dass sich der sorbische und der deutsche Satz so ähnlich wie möglich sind

→ der ideale Satz:

- im Durchschnitt 8 Wörter
- 150 Zeichen
- Interpunktionszeichen, Abkürzungen u. Ä. an gleicher Stelle

3. Ausbau des zweisprachigen Korpus

→ <u>weitere Kriterien für die Bearbeitung:</u>

- nur Sätze (Überschriften, Aufzählungen usw. streichen oder umformulieren)
- jeder Satz endet mit einem Punkt, Ausrufe- oder Fragezeichen
- ein Satz = eine Zeile
- soweit möglich wörtliche Übersetzung → die Sprache darf dabei nicht entstellt und feste Konstruktionen sollten übernommen werden (př.: europaweit = po cyłej Europje)

→ <u>Bearbeitungsprogramm:</u> NOVA Text Aligner

3. Ausbau des zweisprachigen Korpus

File Edit View Left column Right column Block Row Section Bookmarks Info

Untitled (Section 1) 1 / 8

Woni bydlachu tehdy hišće w Radšowje.	Sie wohnten damals noch in Groß Radisch.	1
Zašłu zymu stejach sam w Čelchowje na wikach.	Letzten Winter stand ich alleine in Zschillichau auf dem Markt.	2
Zwjazkowa banka je centralna banka Zwjazkoweje republiki Němskeje.	Die Bundesbank ist die Zentralbank der Bundesrepublik Deutschland.	3
Zwjazkowy statistiski zarjad we Wiesbadenje je wčera nowe ličby wozjewił.	Das Statistische Bundesamt in Wiesbaden hat gestern neue Zahlen veröffentlicht.	4
Krejny ćišć móže přewysoki abo přeniski być.	Der Blutdruck kann zu hoch oder zu niedrig sein.	5
Rozeznawamy periferny a centralny čuwowy system.	Wir unterscheiden das periphere und das zentrale Nervensystem.	6
W tutym padźe móža so hłosy Zwjazka abo wotpowědneho kraja jenož jednotnje wotedać.	In diesem Falle können die Stimmen des Bundes oder des betreffenden Landes nur einheitlich abgegeben werden.	7
Wona zastupuje serbski lud napřećo parlamentam, institucijam a zarjadam.	Sie vertritt das sorbische Volk gegenüber den Parlamenten, Institutionen und Verwaltungen.	8

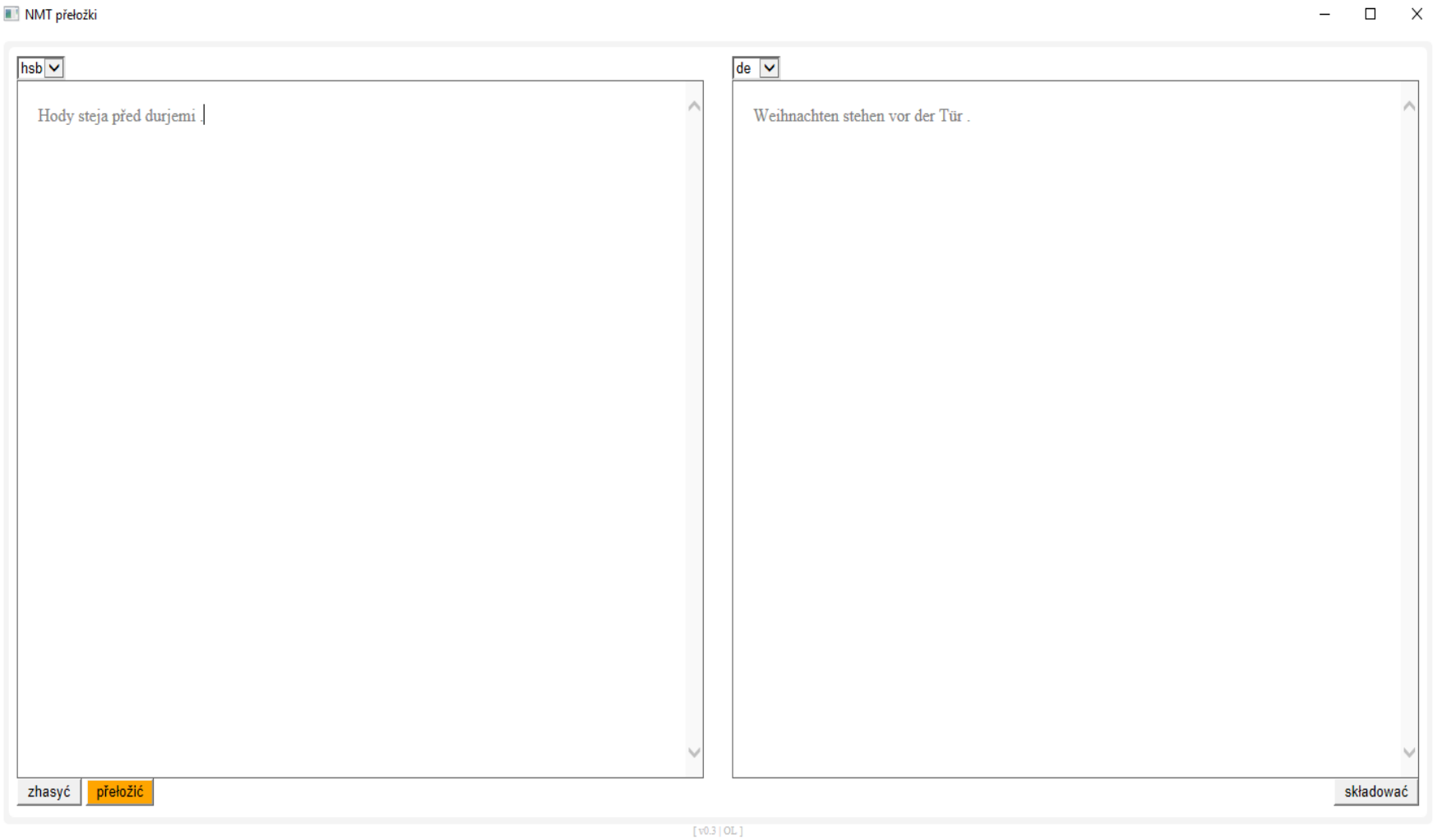
Tools

Block actions: Shift down, Delete, Merge up, First sent. up, Merge down, Last sent. down, Split, Split half

Row actions: Shift down, Delete, Merge up, First sent. up, Merge down, Last sent. down, Split, Split half, Match block lengths after splitting

Highlighting: Mark text:, Mark numbers, Mark if similarity is less then 0.60, Mark if sentence count differs, Mark empty blocks

3. Ausbau des zweisprachigen Korpus



„Man muss es so einrichten, dass einem das Ziel entgegen kommt. “ (Theodor Fontane) …

Wann?
Donnerstag, den 06.02.2020

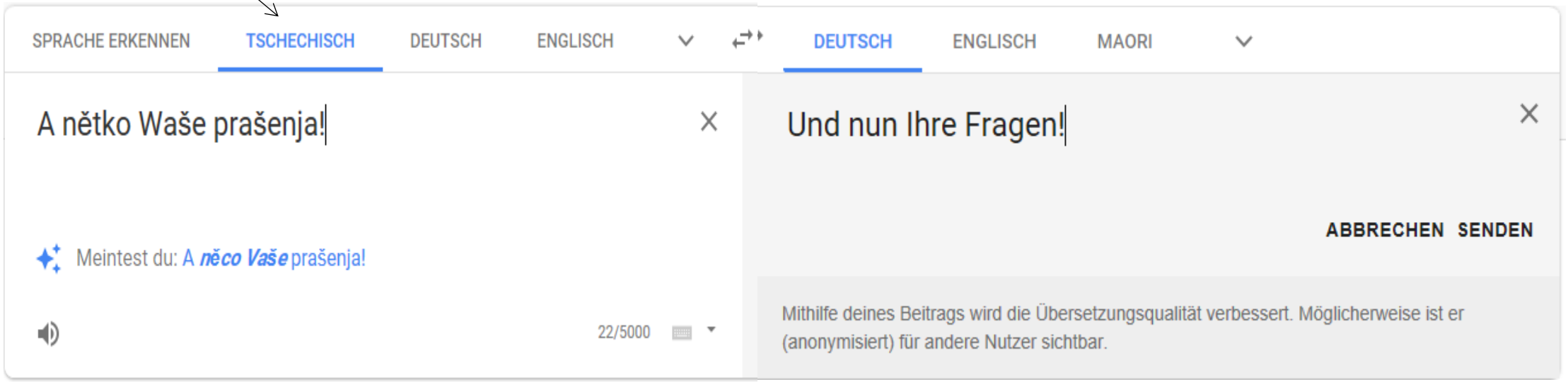
Was?
Sätze formulieren oder bearbeiten

Wer?
Jung und Alt, Fleißige und Interessierte

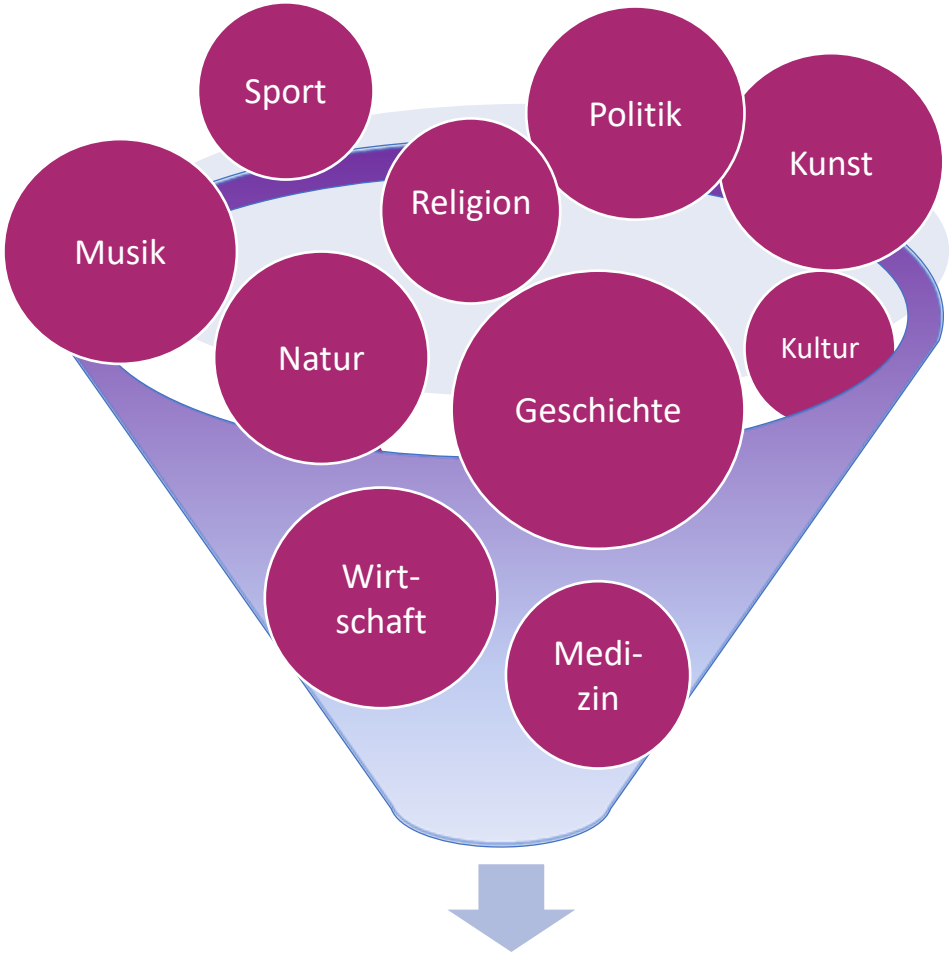
Wo?
Internat des Sorbischen Gymnasiums in Bautzen

... und dieses Ziel sieht so aus:

~~Tschechisch~~ = Obersorbisch → Deutsch (und natürlich umgekehrt)



Du hast Lust und suchst ein sinnvolles Vorhaben für 2020?



Wunderbar, wir brauchen Wörter und Sätze – und suchen DICH!