Global Under-Resourced MEdia Translation (GoURMET)

H2020 Research and Innovation Action
Number: 825299
D1.1 – Survey of relevant low-resource languages

Nature	Report	**Work Package**	WP1
Due Date	30/04/2019	**Submission Date**	30/04/2019
Main authors	Mikel L. Forcada		
Co-authors	Miquel Esplà-Gomis, Juan Antonio Pérez-Ortiz, Víctor M. Sánchez-Cartagena, Felipe Sánchez-Martínez		
Reviewers	Alexandra Birch		
Keywords	survey, languages, resources, machine translation		
Version Control			
v0.8	**Status**	1st Draft	21/04/2019
v1.0	**Status**	Final Version	30/04/2019

Contents

Abstract

This deliverable describes the low-resource languages of interest for project GoURMET, each of which is paired to English in the specific direction (to English or from English), depending on the use cases envisaged by the media partners of the project. A description is provided for each language: a small fact sheet, a study of the main linguistic contrasts with English, an account of corpora and language resources, both monolingual and bilingual, and a brief summary of expected challenges when building machine translation for that language pair.

1 Introduction

This deliverable describes the low-resource languages which are of interest for the GoURMET project. The aim of the deliverable is to inform discussions about which languages to cover in years two and three of the project, as well as to document for researchers the challenges and opportunities related to these language pairs. During the first year of the project GoURMET focuses on Swahili, Gujarati, Bulgarian and Turkish; for years two and three GoURMET will focus on twelve additional languages and on a surprise language.

For each language, this deliverable provides:

- A factsheet for each language, specifying its number of speakers, language family, geographical area, writing system(s), level of standardization (as regards spelling and grammar), dialectal spread or divergence, and similarity to other GoURMET languages or other better-resourced languages.
- A description of important differences or contrasts between the language and English which may have an important effect on machine translation. This will be based in part on a comparison of the values of relevant features (where available or where feasible with the available information) from those collected in the World Atlas of Language Structures (WALS, http://wals.info). Additional literature is cited if used.
- The monolingual and bilingual corpora (generally with English) which are available for them, including:
  - ready-made corpora;
  - monolingual and bilingual text which can be crawled from publicly-available text on the Internet.
- The monolingual and bilingual language resources (generally with English) which are available for them, with an indication of licensing.
- A summary of challenges likely to be faced when building corpus-based (mainly neural) machine translation systems to or from English, as required.

We would like to note that it is, however, difficult to predict in advance how hard it is to build a machine translation system for a given language pair, leaving aside, of course, the effect of the scarcity of data. Researchers in GoURMET addressed this problem (Birch et al., 2008) in the case of *statistical* machine translation for the languages of the European Union; they found that the stronger predictors are (a) the amount of reordering, the (b) morphological complexity of the target language, and (c) the historical relatedness of the two languages. As English is not related to any of the languages covered in this deliverable, the descriptions for each language focuses mainly on the first two predictors.

However, as GoURMET will build *neural* machine translation systems, where target words (a) are sequentially generated from a joint representation of the whole source sentence and (b) may be built from automatically-learned sub-word units, the actual magnitude of the effect of reordering and target-side morphological complexity may be even harder to predict. Descriptions for each case below try to give an account of important linguistic differences between English and each of the languages (with a focus on the specific direction envisaged for each language), as neural

Language	ISO-639 codes	Use case	Partner
Afaan Oromoo	om, orm[2]←	Content creation	BBC
Bosnian	bs, bos←	Content creation	DW
Bulgarian	bg,bul←	Content creation	DW
Croatian	hr,hrv←	Content creation	DW
Gujarati	gu, guj←	Content creation	BBC
Hausa	ha, hau→	Media monitoring	BBC, DW
Igbo	ig ibo←	Content creation	BBC
Korean (North)	ko, kor[3]→	Media monitoring	BBC
Kurdish	ku, kur[4]→	Media monitoring	BBC
Macedonian	mk, mkd (also mac)←	Content creation	DW
Punjabi	pa, pan←	Content creation	BBC
Serbian	sr, srp←	Content creation	BBC, DW
Swahili	sw, swa[5]→	Media monitoring	BBC, DW
Tigrinya	ti, tir←	Content creation	BBC
Turkish	tr, tur←	Content creation	DW
Yoruba	yo, yor←	Content creation	BBC

Table 1: Low-resource languages of interest for GoURMET, with an indication of ISO-639-2 and ISO-639-3 language codes, of whether GoURMET is interested in translations into (←) or from (→) these languages, the use case, and the industrial partner interested.

attention mechanisms would easily deal with one-to-one monotonous translations where little word reordering occurs and the level of morphological complexity is small on both sides.

1.1 Language pairs of interest for GoURMET

The low-resource languages of interest[1] are listed in Table 1. The table indicates the ISO-639 language codes, the use case, whether GoURMET is interested in translations into (←) or from (→) these languages), and the partners interested.

Note that Table 1, and the descriptions provided basically reflect the translation direction and interested partners described in the project proposal. As media partners undertake research for project deliverable D5.2 *Use Case Description and Requirements*, new requirements are revealed. For example, DW have expressed their interest in translation from English to Hausa and Swahili for content creation. The BBC has expressed interest in translation from English into North Korean, Hausa and Swahili for content creation. Additionally the BBC has clarified that for media monitoring, translation from English into Kurdish should focus on the Sorani variety, and expressed

[1] Note that GoURMET has a business news use case which involves better-resourced languages, Spanish and German, which will not be considered in this deliverable.

[2] Inclusive code: there are also gax, Borana–Arsi–Guji–Wallaggaa–Shawaa Oromo; hae, Eastern Oromo; orc, Orma; gaz, West Central Oromo; and ssn, Waata

[3] There is no special code for North Korean

[4] Inclusive code. Includes: ckb, Central Kurdish or Sorani; kmr, Northern Kurdish or Kurmanji; sdh, Southern Kurdish (set of dialects)

[5] Inclusive code. Includes: swc, Congo Swahili; swh, Coastal Swahili; ymk, Makwe; wmw, Mwani

interest on direct translation from Somali into Swahili. Decisions regarding support of new language pairs and/or translation directions will be made as the project progresses.

2 Languages

2.1 Afaan Oromoo (om, orm)

2.1.1 Factsheet

According to Wikipedia, rather than a language, Afaan Oromoo, also called simply *Oromo*, is a group of Afro-Asiatic languages (also called *macrolanguage*) in the Cushitic family; in contrast, Tigrinya (§ 2.14) and Hausa (§ 2.6) are Afro-Asiatic languages in the Chadic family. Afaan Oromoo comprises four main dialects: Southern Oromoo, Eastern Oromoo, Orma, and West-Central Oromoo, and it may be considered a continuum of dialects where the level of mutual understanding depends on the geographical distance between the dialects. It is spoken by about 34,000,000 people[6] in Ethiopia and Kenya (where it is recognized as a minority language) and also in Somalia.

Oromoo is written using the Latin alphabet; long vowels and geminated (doubled) consonants are written by repeating the letter (although this may not be done in a completely consistent way). Stress or tones are not usually marked.

2.1.2 Contrasts with English

The World Atlas of Linguistic Structures describes four different varieties of Afaan Oromoo but it does so with different level of detail. Therefore, the contrasts shown in the following tables may not be completely correct for all of them. Examples are not provided for all contrasts; some of them are taken from Wikipedia, and some from Campbell and King (2010).

[6] Campbell and King (2010) report 24,000,000.

Morphology			
Feature	**Value in English**	**Value in Afaan Oromoo**	**Examples**
Genders	Three, but only in third-person singular pronouns and possessives.	Two genders (semantically or formally assigned)	*ilma* 'son' (masc.); *intala* 'girl, daughter' (fem.); *aduu* 'sun' (fem.).
Morphological case	No case, except for pronouns	Seven cases, marked by suffix or final vowel lengthening[7]	*Ibsaa**n** konkolaataa qaba* 'Ibsaa has a car', lit. 'Ibsaa (+nominative ending) car has'; *mana binne* 'we bought a house', lit. 'house (base case) we bought'; *barumsa afaan**ii*** 'the study of language', lit 'study language ('afaan'+genitive ending); *harkaan* 'by hand, with a hand' ('harka'+instrumental ending), etc.
Noun plural	Always marked (suffix)	Optional (suffix)	*hiriyaa* 'friend'/'friends', *hiriyoota* 'friends'
Adjective plural	Not marked	Suffixing, partial reduplication	*diimaa* 'red' (singular), *diddiimoo* 'red' (plural) (Campbell and King, 2010)
Verb inflection	Tense and (weakly) person	Aspect, mood, gender (3rd person singular), polarity, voice, and person	*beekne* 'we knew' (perfective); *beekte* 'you knew' (perfective); *beekna* 'we know' (imperfective), *beekta* 'you know' (imperfective); *hin beeknu* 'we don't know', *haa beeknu* 'let us know'; *beeki* 'know!' (singular); *beekaa* 'know!' (plural).

[7] Nominative, accusative (base case), dative, ablative, instrumental, locative and genitive.

Function words			
Feature	**Value in English**	**Value in Afaan Oromoo**	**Examples**
Indefinite Articles	Indefinite word distinct from 'one'	No definite or indefinite article	
Definite Articles	Definite word distinct from demonstrative	No definite or indefinite article (Some dialects mark definiteness with a suffix)	*karaa* 'road', *karicha* 'the road'

Syntax			
Feature	**Value in English**	**Value in Afaan Oromoo**	**Examples**
Order of subject, object and verb	Subject–verb–object	Subject–object–verb	*Ibsaan konkolaataa qaba* 'Ibsaa has a car', lit. 'Ibsaa (+nominative ending) car has'
Order of Object, Oblique, and Verb	Verb–Object–Oblique	Object–Oblique–Verb	
Order of Adjective and Noun	Adjective–Noun	Noun–Adjective	
Order of Demonstrative and Noun	Demonstrative–Noun	Noun–Demonstrative	
Order of Numeral and Noun	Numeral–Noun	Noun–Numeral	*nama shan* 'five men', lit. 'man five'
Order of Genitive and Noun	No dominant order	Noun–Genitive	*hojii* 'job', *Caaltuu* (woman's name), *hojii Caaltuu*, 'Caaltuu's job'
Position of Interrogative Phrases in Content Questions	Initial interrogative phrase	Not initial interrogative phrase	
Polar Questions	Interrogative word order	Interrogative intonation only	
Prepositions or postpositions?	Prepositions	Postpositions	
Expression of Pronominal Subjects	Obligatory pronouns in subject position	Subject affixes on verb	*kaleessa dhufne* 'we came yesterday', lit. 'yesterday came-2ND.PL.'

2.1.3 Corpora

Bilingual corpora: The amount of parallel corpora freely available is very scarce. OPUS (http://opus.nlpl.eu/) only contains the Ubuntu v14.10 corpus, which comprises less than 200 sentences.

Monolingual corpora: Oromoo Wikipedia is available at https://om.wikipedia.org/wiki/. As of April 2019 it contains 14,545 entries (see stats on https://stats.wikimedia.org/EN/SummaryOR.

htm). A monolingual dump is periodically made available at https://dumps.wikimedia.org/omwiki/. There is a Wikipedia dump already preprocessed that is available in plain text format (http://hdl.handle.net/11234/1-2735). It contains around 5,000 sentences and 175,000 words.

The Oromo Web Corpus (https://habit-project.eu/wiki/OromoCorpus) contains 5.1 million tokens and is made up of texts collected from the Internet. It can be queried at https://corpora.fi.muni.cz/habit/run.cgi/first_form?corpname=orwac16 but it has not been made available for downloading.

In addition to the BBC (https://www.bbc.com/afaanoromoo), the Voice of America (https://www.voaafaanoromoo.com/) publishes news in Afaan Oromoo.

2.1.4 Resources

Bilingual resources: PanLex contains a bilingual English–Oromo dictionary that can be queried online (https://translate.panlex.org/?langDe=eng-000&langAl=orm-000).

The Glosbe bilingual concordancer (https://glosbe.com/om/en) offers translations into English for Afaan Oromoo words in context.

Monolingual resources: HornMorphoA (https://github.com/hltdi/HornMorpho) is a morphological analyser for languages of the Horn of Africa that supports Oromoo, as well as Tigrinya (§ 2.14).

Wegari (2011) developed a part-of-speech tagger that can be used online (https://nlp.fi.muni.cz/projekty/habit/omtag/index.cgi).

Tesfaye and Abebe (2010) designed an Oromoo stemmer, although it is not available for downloading.

2.1.5 Challenges for corpus-based MT from English

The main challenge when trying to build a machine translation system from English into Afaan Oromoo comes from the scarcity of available parallel corpora. Even if more parallel corpora could be obtained, the following issues would still stand when building an English-to-Oromoo corpus-based machine translation system:

- The order of words in the English and Afaan Oromoo sentence (see contrasts above) is radically different in so many respects (Afaan Oromoo is a subject–object–verb language, uses case suffixes and postpositions instead of prepositions, places demonstratives, numerals and adjectives after nouns, etc.). This property may make the generation of the right word order in Afaan Oromoo difficult.
- Morphological case needs to be marked overtly in Oromo nouns using information which is distributed in English (word order, prepositions, etc.). Similarly, the gender of nouns and adjectives and number of adjectives is not directly present in English words, which makes difficult the generation of grammatically correct Afaan Oromoo.
- Afaan Oromoo is a highly inflected language: for instance, verbs inflect for aspect, mood, gender, polarity, voice and person. This can cause data sparseness problems: if the MT system treats the words as atomic units, the system will likely have to produce words that

have not been observed in the training corpus. It is desirable that that the grammatical suffixes that mark case, number, aspect, mood, etc. are represented as independent tokens to allow the system to generalise better from the training data.

- As Afaan Oromoo is technically not a language but a group of languages or *macrolanguage*, selecting a specific rendition of Afaan Oromoo news which is understood by most literate speakers is no minor challenge.[8]

[8] Partner BBC uses the standard version of the language, as used by the regional state administration, media houses, academic institutions (including the research centers on the language itself); this standard is the one maintained by the Oromia Culture and Tourism Bureau (https://www.oromiatourism.gov.et).

corpus	doc's	sent's	bs tokens	en tokens
OpenSubtitles v2018	17,874	15.2M	97.9M	121.5M
OpenSubtitles v2016	14,014	12.3M	77.9M	97.9M
OpenSubtitles v2011	2,911	2.5M	16.0M	20.3M
OpenSubtitles v2012	2,292	2.1M	13.4M	16.8M
OpenSubtitles v2013	2,303	2.0M	13.2M	16.6M
Tanzil (Quran) v1	30	0.3M	4.7M	5.6M
GNOME v1	488	0.2M	1.1M	1.0M
Ubuntu v14.10	447	0.2M	0.6M	0.8M
total	40,359	34.8M	228M	284M

Table 2: Distribution of the sentences in the OPUS Bosnian–English corpus.

2.2 Bosnian (bs, bos)

2.2.1 Factsheet

The Bosnian language is one of the standardized varieties of the Serbo-Croatian macrolanguage, namely the one mainly used by about 3 million people, living mostly in Bosnia (where it is official) and Herzegovina, but also in Serbia, Montenegro or Kosovo.

Bosnian is written the same Latin script as Croatian (§ 2.4).

2.2.2 Contrasts with English

As regards contrasts with English, Bosnian has essentially the same contrasts as Croatian, see section 2.4.2.

2.2.3 Corpora

Bilingual corpora: The Southeast European Times (SETimes) is a central source of news and information about Southeastern Europe in ten languages: Albanian, Bosnian, Bulgarian, Croatian, English, Greek, Macedonian, Romanian, Serbian and Turkish. The SETimes corpus (http://nlp.ffzg.hr/resources/corpora/setimes/) was compiled and placed in the public domain by Tyers and Serdar Alperen (2010) and refined by the Natural Language Processing group at the University of Zagreb. The Bosnian–English corpus contains approximately 138,000 sentences.

OPUS (http://opus.nlpl.eu) has a Bosnian–English corpus of approximately 35 million sentences, distributed as shown in table 2.

Monolingual corpora: bsWaC is a web corpus collected from the `.ba` top-level domain by Ljubešić and Klubička (2014). It contains 429 million tokens and is annotated with lemma, morphosyntax and dependency syntax layers.

The Twitter corpus BCMS (Ljubešić et al., 2014) contains 379,255,987 words in Bosnian/Croatian/Montenegrin/Serbian. It is distributed as a set of tweet ID's that should be used to rebuild the corpora via the Twitter API.

The Bosnian Wikipedia is available at https://bs.wikipedia.org/wiki/. As of April 2019 it contains 79,223 entries (see stats on https://stats.wikimedia.org/EN/SummaryBS.htm). A monolingual dump is periodically made available at https://dumps.wikimedia.org/bswiki/. There is a Wikipedia dump already preprocessed that is available in plain text format (http://hdl.handle.net/11234/1-2735). It contains around 370,000 sentences and 13 million words.

The W2C (Web to Corpus) corpora is a set of corpora (https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0022-6133-9) for 120 languages automatically collected from Wikipedia and the web. The Bosnian corpus contains around 2 million sentences and 125 million words.

In addition to DW (https://www.dw.com/bs), the following international media outlets produce content in Bosnian: The Voice of America (https://ba.voanews.com/, and TWR360 (https://www.twr360.org/, although mostly multimedia content). These outlets may be interesting sources from which to obtain monolingual corpora.

2.2.4 Resources

Bilingual resources Apertium provides rule-based machine translation between Bosnian/Croatian/Serbian and English. The rule-based system contains bilingual dictionaries and transfer rules released under free licenses (https://github.com/apertium/apertium-hbs-eng).

PanLex contains a bilingual Bosnian–English dictionary that can be queried online (https://translate.panlex.org/?langDe=eng-000&langAl=bos-000).

The Glosbe bilingual concordancer can be used online at https://glosbe.com/en/bs.

The following online machine translation systems support Bosnian–English translation:

- Google Translate (https://translate.google.com)
- Yandex Translate (https://translate.yandex.com)
- Bing Translator (https://www.bing.com/translator)

Monolingual resources Apertium contains a morphological analyser/PoS tagger/morphological generator for Bosnian, Croatian and Serbian (https://github.com/apertium/apertium-hbs) with 58,004 stems.

2.2.5 Challenges for corpus-based MT from English

The linguistic contrasts between English and Bosnian are basically the same as those between English and Croatian (§ 2.4.2), and so are the challenges for corpus-based MT from English (§ 2.4.5). The main additional problem may lie in harvesting new corpora for Bosnian, as language identification is very likely to (expectedly) classify Croatian, or even Serbian text as Bosnian or vice-versa.

2.3 Bulgarian (bg, bul)

2.3.1 Factsheet

Bulgarian is an Indo-European, Slavic language, as are the languages in the Serbo-Croatian macrolanguage (such as Bosnian (§ 2.2), Croatian (§ 2.4), and Serbian (§ 2.12)), and the closely-related Macedonian (§ 2.10). According to Wikipedia, it is spoken by 8–9 million people, mainly in Bulgaria (where it is official) but also in Albania, the Czech Republic, Hungary, Moldova, Romania, Serbia, and the Ukraine (where it has varying levels of recognition as a minority language). It is one of the 24 official languages of the European Union.

Bulgarian and Macedonian form the East South Slavic group, which departs dramatically from other Slavic languages in that they have done away with the case declension system, do not have a verb infinitive, and have acquired a definite article (written as a suffix), but have a much richer verb system, including evidentiality (for instance, it morphologically marks whether the action was witnessed directly by the speaker or was reported to them).

Bulgarian uses the Cyrillic alphabet.

2.3.2 Contrasts with English

Examples are given in transliteration.

Function words			
Feature	**Value in English**	**Value in Bulgarian**	**Examples**
Definite articles	Definite word distinct from demonstrative, separate word	Definite word distinct from demonstrative, suffix.	*kniga* (‘book’), *knigata*, ‘the book’.
Indefinite articles	Indefinite word distinct from *one*	No indefinite article	*kniga* (‘book’, ‘a book’).

Morphology			
Feature	**Value in English**	**Value in Bulgarian**	**Examples**
Coding of evidentiality	No grammatical evidentials	Evidentials part of the verb system	*Toĭ e bil* (‘He was’, direct evidence) vs. *Toĭ bil* (‘He reportedly was’, indirect evidence).
The perfect (verbs)	Uses *have* as an auxiliary	Does not use *have*	*Az sŭm napisal* (‘I have written’, where *Az sŭm* means ‘I am’).
2nd person morphological imperative	Not different from base form	Special and different for singular and plural	*karaĭ!* (‘Drive!’, sing.); *karaĭte!* (‘drive’, plural).
Perfective vs. Imperfective Aspect:	No grammatical marking	Grammatical marking	*Kogato* ***cheta*** *kniga* (‘When I read a book’, imperfective, unfinished action); *Kogato* ***procheta*** *knigata* (‘When I finish reading the book’, perfective, finished action).

Syntax			
Feature	**Value in English**	**Value in Bulgarian**	**Examples**
Order of subject and verb	Subject–Verb	No preference	*Pozdravi Ivan momichetata* or *Ivan pozdravi momichetata* (‘Ivan greets the girls’)
Polar (yes/no) questions	No question particle, signalled by word order or auxiliary	Particle in the second position	*Imate kniga* (‘You have got a book’), *Imate li kniga?* (‘Have you got a book?’)
Expression of pronominal subjects	Mandatory pronoun in subject position	Person affix in verb	*Vidyakhte* (‘You saw’), *vidyakhme* (‘We saw’), *Vidyakha* (‘They saw’).
’Want’ Complement Subjects	Subject is left implicit	Subject is expressed overtly	*Iskam da pluvam* (‘I want to swim’, lit. ‘I want that I swim’, parallel to *Iskam da pluvash* ‘I want you to swim’).

2.3.3 Corpora

Bilingual corpora: The Southeast European Times (SETimes) is a central source of news and information about Southeastern Europe in ten languages: Albanian, Bosnian, Bulgarian, Croatian, English, Greek, Macedonian, Romanian, Serbian and Turkish. The SETimes corpus (http://nlp.ffzg.hr/resources/corpora/setimes/) was compiled and put in the public domain by Tyers and Serdar Alperen (2010) and refined by the Natural Language Processing group at the University of Zagreb. The Bulgarian–English corpus contains approximately 213,000 sentences.

ParaSol (http://parasolcorpus.org/) is a parallel aligned corpus of translated and original belletristic texts in Slavic and some other languages. The amount of parallel corpora depends on the particular language pair. Languages include Bulgarian, Belarusian, Czech, Croatian, Macedonian, Polish, Russian, Slovak, Slovene, Serbian, Ukrainian, Upper Sorbian, German, English, Dutch, Spanish, French, Italian, and a few others. Croatian texts are tagged and lemmatized. The Bulgarian part has 2,125,936 tokens and 47,172 lemmas, whereas the English part has 814,289 tokens and 19,886 lemmas. Access to ParaSol and downloads are provided by a web interface which requires authentication.

Europarl (http://www.statmt.org/europarl/) has a Bulgarian–English section with around 400,000 parallel sentences.

The DGT translation memory (https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory) contains 4,316,876 Bulgarian–English translation units, while the EAC translation memory (https://ec.europa.eu/jrc/en/language-technologies/eac-translation-memory) contains only 4,061 translation units and the ECDC translation memory (https://ec.europa.eu/jrc/en/language-technologies/ecdc-translation-memory) contains 2,567 translation units.

The Digital Corpus of the European Parliament (https://ec.europa.eu/jrc/en/language-technologies/dcep) contains the majority of the documents published on the European Parliament's official website. The Bulgarian–English section contains around 35 million words in Bulgarian. The JRC-Acquis corpus (https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis) is a parallel corpus extracted from the body of European Union (EU) law applicable in the the EU Member States. It contains around 46,000 Bulgarian–English parallel sentences.

In addition to the aforementioned sentence-aligned parallel corpora published by European institutions, DGT-Acquis (https://ec.europa.eu/jrc/en/language-technologies/dgt-acquis) is a paragraph-aligned corpus that contains around 56 million words in Bulgarian. It needs to be sentence-aligned before it can be used for training MT systems.

The release 4.0 of the Paracrawl parallel corpora collection (lhttps://paracrawl.eu/releases.html) contains an Bulgarian–English parallel corpus with around 1 million sentences.

The EUR-Lex Corpus (https://www.sketchengine.eu/eurlex-corpus/) is a multilingual corpus in all the official languages of the European Union. The corpus has been built from HTML files available in the EUR-Lex database. It is released under CC-BY-NC-SA licence. It is paragraph-aligned and contains around 17 million Bulgarian–English paragraph pairs.

The EMP-BTB-CSLI-MWA and EMP-BTB-JH0-MWA datasets (http://bultreebank.org/bg/btb-results-from-euromatrixplus-project/) were produced during the EuroMatrixPlus Project. They respectively contain 893 and 250 parallel sentences manually aligned at the word level.

The novel "1984" by George Orwell tagged with lemma and part of speech in Bulgarian and English can be downloaded from https://www.clarin.si/repository/xmlui/handle/11356/1043. English original has 79,718 sentences and 106,4424 words. The corpus is licensed under a CC BY-NC-SA 4.0 license.

The QTLeap parallel corpus, available via Portulan Clarin (https://portulanclarin.net/), is composed by 4,000 question and answer pairs in the domain of computer and IT troubleshooting.

As part of the Machine Translation of IT domain shared task at the WMT16 conference (http://www.statmt.org/wmt16/it-translation-task.html) a parallel corpus built from software localization files was released. It contains around 100,000 Bulgarian–English parallel sentences.

corpus	doc's	sent's	bg tokens	en tokens
OpenSubtitles v2018	52,151	42.9M	289.1M	349.9M
OpenSubtitles v2016	42,081	35.4M	236.3M	287.8M
OpenSubtitles v2012	23,580	20.5M	136.1M	168.0M
OpenSubtitles v2013	19,324	17.8M	117.9M	146.6M
OpenSubtitles v2011	19,420	16.7M	111.0M	137.6M
EMEA v3	1,596	1.1M	14.6M	11.0M
EUbookshop v2	740	0.2M	10.0M	12.8M
Tanzil (Quran) v1	15	0.1M	2.3M	2.8M
GNOME v1	1,326	0.6M	2.4M	2.6M
Tatoeba v2	1	11.2k	92.5k	3.6M
OpenSubtitles v1	209	0.2M	1.6M	2.0M
Wikipedia v1.0	1	79.8k	1.4M	1.7M
KDE4 v2	651	0.1M	0.7M	0.8M
Ubuntu v14.10	350	79.0k	0.3M	0.5M
GlobalVoices v2017q3	293	5.7k	0.1M	0.1M
GlobalVoices v2015	262	4.6k	0.1M	0.1M
total	162,000	136M	924M	1138M

Table 3: Distribution of the sentences in the Opus Bulgarian–English corpus.

Opus (http://opus.nlpl.eu) has a Bulgarian–English corpus of approximately 136 million sentences distributed as shown in table 3.

Monolingual corpora: Bulgarian Wikipedia is available at https://bg.wikipedia.org/wiki/. As of April 2019 it contains 248,310 entries (see stats on https://stats.wikimedia.org/EN/SummaryBG.htm). A monolingual dump is periodically made available at https://dumps.wikimedia.org/bgwiki/. There is a Wikipedia dump already preprocessed that is available in plain text format (http://hdl.handle.net/11234/1-2735). It contains around 2.1 million sentences and 37 million words.

The W2C (Web to Corpus) corpora is a set of corpora (http://hdl.handle.net/11858/00-097C-0000-0022-6133-9) for 120 languages automatically collected from Wikipedia and the web. The Bulgarian corpus contains around 1.4 million sentences and 50 million words.

Crawls of Bulgarian news text are available at http://data.statmt.org/news-crawl/bg/. As of April 2019, the crawls contain 35 million sentences and 600 million words.

The bgTenTen corpus (https://www.sketchengine.eu/bgtenten-bulgarian-corpus/) is a Bulgarian corpus made up of texts collected from the Internet. It is available at the Sketch Engine platform, but it requires a subscription.

The Brown corpus of Bulgarian (http://dcl.bas.bg/Corpus/home_en.html) consists on 500 text samples distributed in 15 categories from two types of texts: fiction and informative prose. The corpus amounts to 1,001,286 words.

The Bulgarian National Corpus (https://www.sketchengine.eu/bulgarian-national-corpus/) is a Bulgarian corpus made up of texts collected from various sources such as scanned books, transcribed data, Internet texts, etc. It contains 419 million words. It is available at the Sketch Engine

platform, but it requires a subscription.

The BgSpeech database (http://bgspeech.net/en/resources_en.html) contains transcriptions of spoken Bulgarian, which are available for research and academic purposes.

The C4Corpus (https://dkpro.github.io/dkpro-c4corpus/) is extracted from the on-line available CommonCrawl, a massive crawl of documents from the Internet and contains monolingual text in Bulgarian.

The BulTreeBank corpus (Simov et al., 2002) is a syntactically annotated Bulgarian corpus. A subset of it can be downloaded at http://bultreebank.org/en/resources/. The same website also contains some sections of the Bulgarian National Reference Corpus (available for research purposes).

In addition to DW (https://www.dw.com/bg), the following international media outlets produce content in Bulgarian: China Plus (formerly China Radio International, http://bulgarian.cri.cn/) Vatican Radio (https://www.vaticannews.va/bg.html), and TWR360 (https://www.twr360.org/, although most of it is multimedia content). These outlets may be interesting sources from which to obtain monolingual corpora.

2.3.4 Resources

Monolingual resources: Apertium has fair-coverage morphological resources for Bulgarian (https://github.com/apertium/apertium-bul, 8578 lemmata).

Besides Apertium, the MULTEXT-East lexicons (http://hdl.handle.net/11356/1041@format=cmdi) also contain morphological inflection information for Bulgarian words.

Moreover, there are PoS tagging models for 3 different tools available at http://bultreebank.org/en/resources/part-speech-tagging-bultreebank-bulgarian-taggers/.

There are monolingual resources for syntactic analysis of Bulgarian too, such as the BURGER resource grammar (http://bultreebank.org/en/bulgarian-resource-grammar-burger/) and a dependency parser trained on the BulTreeBank corpus (Marinov and Nivre, 2005).

Bilingual resources: PanLex contains a bilingual Bulgarian–English dictionary that can be queried online (https://translate.panlex.org/?langDe=eng-000&langAl=bul-000).

A crowd-sourced Bulgarian–English dictionary (http://www.seas.upenn.edu/~nlp/resources/TACL-data-release/dictionaries.tar.gz) has been made available by Pavlick et al. (2014).

Open Multilingual Wordnet (http://compling.hss.ntu.edu.sg/omw/) contains synsets in Bulgarian linked to the corresponding entry in the Princeton Wordnet of English (https://wordnet.princeton.edu/). A Bulgarian–English bilingual dictionary could be easily extracted from this resource.

The Glosbe bilingual concordancer can be used online at https://glosbe.com/en/bg.

The following online machine translation systems support Bulgarian–English translation:

- Google Translate (https://translate.google.com)
- Yandex Translate (https://translate.yandex.com)
- Bing Translator (https://www.bing.com/translator)

2.3.5 Challenges for corpus-based MT from English

Since Bulgarian is an official language of the EU, bilingual corpus scarcity is less of an issue for the Bulgarian–English language pair than it is for most of the languages in GoURMET.

These are the main divergences with English when generating Bulgarian:

- Definiteness needs to be marked overtly in Bulgarian nouns using information from the determiners in English.
- Pronominal subjects need to be expressed as verb affixes in Bulgarian using information from the personal pronouns in English.
- Subject of the complement of the verb *want* needs to be expressed overtly in Bulgarian using information from the subject of the verb *want*.
- Evidentiality, perfective/imperfective aspect, and number of 2nd person morphological imperative, which are not explicitly encoded in English, need to be marked in Bulgarian verbs. The MT system would need to make use of all the information in the SL sentence (and maybe in the surrounding ones) to guess the value of these morphological features.
- The polarity of questions needs to be expressed with a particle by rearranging the word order (including a possible auxiliary verb) in English.
- Bulgarian is a highly inflected language. This can cause data sparseness problems if the MT system treats the words as atomic units. It is desirable that that the grammatical suffixes that mark defineteness, evidentiality, etc. are represented as independent tokens to allow the system to generalise better from the training data.

2.4 Croatian (hr, hrv)

2.4.1 Factsheet

According to Wikipedia, Croatian is the standardized variety of the Serbo-Croatian macro-language used by about 5.6 million Croats, principally in Croatia (where it is official), Bosnia and Herzegovina (where it is one of the official language), the Serbian province of Vojvodina (where it is recognized as a minority language), and other neighboring countries. It is one of the 24 official languages of the European Union. Bosnian (§ 2.2) and Serbian (§ 2.12) are the other two languages in this projects belonging to the same macro-language. The Serbo-Croatian macro-language, as well as Bulgarian (§ 2.3) and Macedonian (§ 2.10) in this project, belong to the Slavic group inside the Indo-European family.

Croatian is written with the Latin script, supplemented with the following letters: *č*, *ć*, *đ*, *š*, and *ž* (upper case *Č*, *Ć*, *Đ*, *Š*, and *Ž*).

2.4.2 Contrasts with English

Function words			
Feature	**Value in English**	**Value in Croatian**	**Examples**
Definite articles	Definite word distinct from demonstrative	No definite or indefinite article	*knjiga* (‘book’, ‘the book’, ‘a book’).
Indefinite articles	Indefinite word distinct from *one*	No definite or indefinite article	*knjiga* (‘book’, ‘the book’, ‘a book’).

Morphology			
Feature	**Value in English**	**Value in Croatian**	**Examples**
Number of cases	2 cases, only for pronouns and similar words	5–7 cases	***Knjiga*** *je tamo* (‘**the book** is there’, nominative); *Imam* ***knjigu*** (‘I have **a book**’, accusative); *Cijena* ***knjige*** (‘The price **of the book**’, genitive); *Govorio je o* ***knjizi*** (‘He talked about **the book**’, prepositional); *Došao je s* ***knjigom*** (‘He arrived with **the book**’, instrumental).
Position of case affixes	No case affixes or adpositional clitics	Case suffixes	*knjig****a*** (‘book’, nominative), *knjig****om*** (‘book’, instrumental)
The morphological Imperative	No second-person imperatives	Second singular and second plural	*Pročitati* (‘read’); *pročitaj!* (‘read!’, singular); *pročitajte!* (‘read!’, plural).

Syntax			
Feature	**Value in English**	**Value in Croatian**	**Examples**
Yes–no questions	No question particle, uses word order or auxiliary	Question particle in a few specific positions	*Razumijete.* ('You understand'); *Da li razumijete?* or *Razumijete li?* ('Do you understand?').
Expression of pronominal subjects	Obligatory pronouns in subject position	Subject affixes on verbs, number and gender agreement on participles.	*Razumije* ('He or she understands'); *Razumju* ('They understand'); *Razumio je* ('He has understood'), *Razumjela je* ('She has arrived'), *Razumjeli su* ('They [masc] have arrived'), etc.
Multiple negation	Changes polarity	Does not change polarity, often mandatory	*Nigdje nisam vidio nikoga* ('I haven't seen anyone anywhere', lit. 'Nowhere I haven't seen no one'); *Nisam vidio nikoga* ('I haven't seen anyone', lit. 'I haven't seen no one').
Order of genitive and noun	No dominant order	Noun–genitive but some constructs differ	*Cijena knjige* 'The price (*cijena*) of the book (*knjige*)'; *Auto moje majke* 'My mother's car'; but: *Mamin auto* 'Mom's car'; *Čovjekov auto* 'The man's car' (with adjective-forming suffixes *-in* and *-ov* applied to *mama* and *čovjek*)

2.4.3 Corpora

Bilingual corpora: The Southeast European Times (SETimes) is a central source of news and information about Southeastern Europe in ten languages: Albanian, Bosnian, Bulgarian, Croatian, English, Greek, Macedonian, Romanian, Serbian and Turkish. The SETimes corpus (http://nlp.ffzg.hr/resources/corpora/setimes/) was compiled and put in the public domain by Tyers and Serdar Alperen (2010) and refined by the Natural Language Processing group at the University of Zagreb. The Croatian–English corpus contains approximately 200,000 sentences.

ParaSol (http://parasolcorpus.org/) is a parallel aligned corpus of translated and original belletristic texts in Slavic and some other languages. The amount of parallel corpora depends on the particular language pair. Languages include Bulgarian, Belarusian, Czech, Croatian, Macedonian, Polish, Russian, Slovak, Slovene, Serbian, Ukrainian, Upper Sorbian, German, English, Dutch, Spanish, French, Italian, and a few others. Croatian texts are tagged and lemmatized. The Croatian part has 1,184,904 tokens and 55,538 lemmas, whereas the English part has 814,289 tokens and 19,886 lemmas. Access to ParaSol and downloads are provided by a web interface which requires authentication.

The hrenWaC — Croatian–English Parallel Web Corpus http://nlp.ffzg.hr/resources/corpora/hrenwac/ consists of 99,001 Croatian–English sentence pairs. It is published under the CC-BY-SA license.

The Croatian–English TED talks parallel corpus (http://nlp.ffzg.hr/resources/corpora/ted-talks/) is a collection of parallel sentences extracted from the Croatian-English TED talks transcripts

corpus	doc's	sent's	en tokens	hr tokens
OpenSubtitles v2018	46,239	37.5M	305.7M	243.4M
OpenSubtitles v2016	34,580	28.8M	234.2M	185.0M
OpenSubtitles v2013	18,640	16.6M	136.9M	105.6M
OpenSubtitles v2012	18,002	16.4M	132.2M	102.2M
OpenSubtitles v2011	11,254	10.2M	82.5M	64.5M
GNOME v1	886	0.3M	1.9M	1.0M
TedTalks v1	1	86.3k	1.5M	1.3M
KDE4 v2	809	0.1M	0.8M	0.5M
Ubuntu v14.10	293	52.7k	0.5M	0.2M
EUbookshop v2	23	6.2k	0.2M	0.2M
total	130,727	109.9M	896M	704M

Table 4: Distribution of the sentences in the Opus Croatian–English corpus.

available from WIT3. The corpus consists of 86.348 sentence pairs, 2.384.887 tokens.

The Tourism Croatian–English Parallel Corpus (Esplà-Gomis et al., 2014) is a sentence aligned parallel corpus built by automatically crawling 25 websites from the tourism domain. It contains 87,024 aligned segments.

The release 4.0 of the Paracrawl parallel corpora collection (lhttps://paracrawl.eu/releases.html) contains an Croatian–English parallel corpus with 1 million sentences.

The DGT translation memory (https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory) contains 2,288,146 Croatian–English translation units, while the EAC translation memory (https://ec.europa.eu/jrc/en/language-technologies/eac-translation-memory) contains only 573 translation units.

Opus (http://opus.nlpl.eu) has a Croatian–English corpus of approximately 110 million sentences distributed as shown in table 4.

Monolingual corpora: hrWaC is a web corpus collected from the .hr top-level domain by Ljubešić and Klubička (2014). It contains 1,900 million tokens and is annotated with lemma, morphosyntax and dependency syntax layers.

The Twitter corpus of BCMS (Ljubešić et al., 2014) contains 379,255,987 words in Bosnian/Croatian/Montenegrin/Serbian. It is distributed as a set of tweet ids that should be used to rebuild the corpora via the Twitter API.

The Croatian Wikipedia is available at https://hr.wikipedia.org/wiki/. As of April 2019 it contains 201,390 entries (see stats on https://stats.wikimedia.org/EN/SummaryHR.htm). A monolingual dump is periodically made available at https://dumps.wikimedia.org/hrwiki/. There is a Wikipedia dump already preprocessed that is available in plain text format (http://hdl.handle.net/11234/1-2735). It contains around 2.1 million sentences and 37 million words.

The W2C (Web to Corpus) corpora is a set of corpora (http://hdl.handle.net/11858/00-097C-0000-0022-6133-9) for 120 languages automatically collected from Wikipedia and the web. The Croatian corpus contains around 1.8 million sentences and 100 million words.

In addition to BBC (who have an archived Croatian news page, http://www.bbc.co.uk/croatian/,

but no current page) and DW (https://www.dw.com/hr/), the following international media outlets produce content in Croatian: Vatican News (https://www.vaticannews.va/hr.html), China Plus (formerly China Radio International, http://croatian.cri.cn/) and TWR360 (https://www.twr360.org/, although mostly multimedia content). These outlets may be interesting sources from which to obtain monolingual corpora.

2.4.4 Resources

Bilingual resources: Apertium provides rule-based machine translation between Bosnian/Croatian/Serbian and English. The rule-based system contains bilingual dictionaries and transfer rules released under free licenses (https://github.com/apertium/apertium-hbs-eng).

PanLex contains a bilingual Croatian–English dictionary that can be queried online (https://translate.panlex.org/?langDe=eng-000&langAl=hrv-000).

A crowd-sourced Croatian–English dictionary (http://www.seas.upenn.edu/~nlp/resources/TACL-data-release/dictionaries.tar.gz) has been made available by Pavlick et al. (2014).

The set of open-source bilingual dictionaries FreeDict (https://github.com/freedict/fd-dictionaries) contains an Croatian–English bilingual dictionary.

Open Multilingual Wordnet (http://compling.hss.ntu.edu.sg/omw/) contains synsets in Croatian linked to the corresponding entry in the Princeton Wordnet of English (https://wordnet.princeton.edu/). A Croatian–English bilingual dictionary could be easily extracted from this resource.

The Glosbe bilingual concordancer can be used online at https://glosbe.com/en/hr.

The following online machine translation systems support Croatian–English translation:

- Google Translate (https://translate.google.com)
- Yandex Translate (https://translate.yandex.com)
- Bing Translator (https://www.bing.com/translator)

Monolingual resources Apertium contains a morphological analyser/PoS tagger/morphological generator for Bosnian, Croatian and Serbian (https://github.com/apertium/apertium-hbs) with 58,004 stems.

The hrLex morphological lexicon (Ljubešić et al., 2016), released under a free licence, can be used together with a CRF tagger (Ljubešić and Erjavec, 2016) to morphologically analyze Croatian text.

2.4.5 Challenges for corpus-based MT from English

Croatian is an official language of the EU since Croatia joined in July 2013; as a result, it is better resourced than most of the languages of interest for GoURMET. However, it may not be so well-resourced if one considers the specific domain of GoURMET, that is, news, as most of the EU-official material will likely be out-of-domain.

These are the main challenges when generating Croatian:

- Morphological case needs to be marked overtly in Croatian nouns and other noun-phrase elements such as adjectives and determiners, using information which is distributed in English (word order, prepositions, etc.).
- Due to the fact that nouns in Croatian inflect for up to 7 different grammatical cases, data sparseness issues can arise if the machine translation system operates on whole Croatian words. The inflected nouns that need to be generated to correctly translate an English noun may have not been observed in the training corpus. In fact, Klubička et al. (2018) showed that agreement errors are reduced when NMT operating on byte-pair-encoded (Sennrich et al., 2016) words is used instead of SMT (operating on whole words).
- One needs to generate multiple negations in negative sentences.
- The order of genitive constructions, which is free in English, needs to be set in Croatian depending on set of factors such as whether the phrase that acts as the possessor is a proper noun or not.

2.5 Gujarati (gu, guj)

2.5.1 Factsheet

According to Wikipedia, Gujarati has around 55,000,000 speakers. It is spoken (and official) in the state of Gujarat and in two *Union territories* (Daman and Diu and Dadra and Nagar Haveli) (India). It belongs to a very large family, the Indo-European family, as do many other languages in this project, namely Bosnian (§ 2.2), Croatian (§ 2.4), Serbian (§ 2.12), Bulgarian (§ 2.3), and Macedonian (§ 2.3); however, these five languages are Slavic, while Gujarati, Punjabi (§ 2.11) and Kurdish (§ 2.9), also in this project, belong to the Indo-Iranian group.

Gujarati is written in a script of its own, the Gujarati script, which, as the Devanagari script used for Hindi, is an *alphasyllabary* in which consonant–vowel groups are written as a single character, with vowels being secondary to consonants.

2.5.2 Contrasts with English

Gujarati examples are given in transliteration.

Syntax			
Feature	**Value in English**	**Value in Gujarati**	**Examples**
Order of Subject, Object and Verb	Subject–Object–Verb	Subject–Verb–Object	*Chōkarō gāya ju'ē chē* ('The boy sees the cow', lit. 'Boy cow sees')
Adpositions: prepositions or postpositions?	Prepositions	Postpositions	*ghara* ('The house'), *gharamāṁ* ('In the house'); *kāra* ('The car'). *kāra māṁ* ('In the car').
Position of Interrogative Phrases in Content Questions	Initial interrogative phrase	Not initial interrogative phrase	*Tē **ghara** ju'ē chē* ('She sees **the house**') *Tē **śuṁ** ju'ē chē?* ('**What** does she see?)
Polar questions	Interrogative word order	Same order as affirmative	*Tē ghara ju'ē chē* ('She sees the house') *Śuṁ tē ghara ju'ē chē?* ('Does she see the house?')

Morphology			
Feature	**Value in English**	**Value in Gujarati**	**Examples**
Number of genders	Three, but only in 3rd person singular pronouns and possessives	Two, masculine and feminine	*Chōkarō ūñcō chē* ('The boy is tall', masc.) *Chōkarī̃ ūñcī̃ chē* ('The girl is tall', fem.)
The transitive verb agrees with...	...the A argument ('agent'), if it does	...the A argument in imperfective verb forms, and the P argument ('patient') in perfective verb forms (split ergativity[9])	The verb agrees in number with P (and not with A) in perfective constructs such as *Ēka mātā'ē ēka bāl̇akanē jōyō chē* ('One mother-ERG one child-ACC has-seen-SING'); *Ēka mātā'ē bē bāl̇akōnē jōyā chē* ('One mother has seen two children', lit. 'One mother-ERG two children-ACC has-seen-PLU'); *Bē mātā'ō'ē ēka bāl̇akanē jōyō chē* ('Two mothers see one child', lit. Two mothers-ERG one child-ACC has-seen-SING') but it agrees with A in imperfective constructs as it does in English.
Plurality in nouns	Always expressed	Optional	
Morphological case	No case (except for pronouns, etc.)	Three cases: nominative, oblique, and, for non-feminine nouns, locative; pronouns have more morphological cases	

2.5.3 Corpora

Monolingual corpora: A crawl of Gujarati news text is available at http://data.statmt.org/news-crawl/gu/. The monolingual dump of the Gujarati wikipedia is periodically made available at https://dumps.wikimedia.org/guwiki/. The C4Corpus (https://dkpro.github.io/dkpro-c4corpus/) is extracted from the on-line available CommonCrawl, a massive crawl of documents from the

[9] A grammatical process shows *ergative* alignment when the subject of an intransitive verb (S) identifies with the object of a transitive verb (P, *patient*). It shows *accusative* alignment when the subject of a transitive verb (A, *agent*) identifies with the subject of an intransitive verb (S). Split ergativity (coexistence of ergative and accusative alignments) occurs only in some contexts, for instance when verbs are perfective in the case of Gujarati. Punjabi (§ 2.11) also shows split ergativity.

Internet. Another Gujarati news corpus crawled from the Internet in 2014 is made available by the University of Leipzig (https://clarinws.informatik.uni-leipzig.de/clarinwebservices/sentences/11022/0000-0000-7F62-4/sentencestext/). A set of corpora for 120 languages automatically collected from Wikipedia and the web is released under the name of W2C (Web to Corpus) corpora (http://hdl.handle.net/11858/00-097C-0000-0022-6133-9). A collection of monolingual corpora are made available by the Leipzig University through their corpora portal (http://wortschatz.uni-leipzig.de/en/download/); custom tools need to be used to download the corpora.

The Deltacorpus 1.1 (http://hdl.handle.net/11234/1-1743) contains texts in 107 languages from the W2C corpus (http://hdl.handle.net/11858/00-097C-0000-0022-6133-9). The first 1,000,000 tokens of each language are part-of-speech-tagged by the de-lexicalized tagger described in Yu et al. (2016, LREC, Portorož, Slovenia)

There is a monolingual Gujarati corpus released as part of the Indian Languages Corpora Initiative phase–II. It contains about 30,000 sentences of general domain. Sentences have been part-of-speech-tagged according to BIS (Bureau of Indian Standards) tagset. It is released under a research license and available under request at http://tdil-dc.in/index.php?option=com_download&task=showresourceDetails&toolid=1882&lang=en.

The Gujarati News Corpus (SRIMCA), a corpus created at the *Shrimad Rajchandra Institute of Management and Computer Application*, covers news articles from different newspapers published in Gujarati. It contains 156, 210, 101 and 50 news articles in the domain of business, crime, politics and sports, respectively. It is released under a research license and is available under request at http://tdil-dc.in/index.php?option=com_download&task=showresourceDetails&toolid=1998&lang=en.

The EMILLE Lancaster Monolingual Corpus (http://catalog.elra.info/en-us/repository/browse/ELRA-W0038/) provides monolingual data in seven South Asian languages (about 58,880,000 tokens), including Gujarati. This corpus is only available after paying a fee.

In addition to the BBC, only TWR360 has been identified as a media outlet containing Gujarati content (https://www.twr360.org/, although most of it is multimedia content). These outlets may be interesting sources from which to obtain monolingual corpora.

Bilingual corpora: The EMILLE Lancaster Corpus (http://catalog.elra.info/en-us/repository/browse/ELRA-W0038/) provides parallel sentences between English (about 200,000 tokens) and seven Sout Asian languages, including Gujarati. This corpus is only available after paying a fee.

Some parallel corpora are available for Gujarati–English language pair. Three of them can be downloaded from the OPUS corpora collection. In this case, all of them are obtained from three different free/open-source software projects (details are available in table 5):

- GNOME: http://opus.nlpl.eu/download.php?f=GNOME/v1/moses/en-gu.txt.zip
- Ubuntu: http://opus.nlpl.eu/download.php?f=Ubuntu/v14.10/moses/en-gu.txt.zip
- KDE: http://opus.nlpl.eu/download.php?f=KDE4/v2/moses/en-gu.txt.zip

Some bilingual corpora have been provided by the organizers of the WMT 2019 news translation task:

- Wikipedia titles: http://data.statmt.org/wikititles/v1/wikititles-v1.gu-en.tsv.gz (12,000 titles)

corpus	doc's	sent's	en tokens	gu tokens
GNOME v1	1145	0.5M	2.4M	4.4M
KDE4 v2	210	57.5k	0.3M	0.4M
Ubuntu v14.10	230	34.7k	0.2M	0.2M
total	1585	592.2k	2.9M	5.0M

Table 5: Distribution of the sentences in the Opus Gujarati–English corpus.

- A Bible corpus: http://data.statmt.org/wmt19/translation-task/bible.gu-en.tsv.gz (7,800 sentence pairs)
- The software localization corpora in OPUS: http://data.statmt.org/wmt19/translation-task/opus.gu-en.tsv.gz (108,000 sentence pairs)
- A parallel corpus extracted from Wikipedia and contributed by Alexander Molchanov (Yandex): http://data.statmt.org/wmt19/translation-task/wikipedia.gu-en.tsv.gz (18,000 sentence pairs)
- A corpus crawled from the Internet by WMT organizers. The clean version, http://data.statmt.org/wmt19/translation-task/govin-clean.gu-en.tsv.gz, contains 11,000 sentence pairs. The raw version http://data.statmt.org/wmt19/translation-task/govin-raw.gu-en.tsv.gz is much larger.
- An English–Gujarati from health domain developed by the *English to Indian Language Machine Translation (EILMT) Consortium*. The size of the corpus is 14,961 sentence pairs and is available under request at https://tdil-dc.in/index.php?option=com_download&task=showresourceDetails&toolid=1785&lang=en.

2.5.4 Resources

Monolingual resources: A number of natural-language-processing tools are available for Gujarati. The `indicNLP` (https://github.com/nisargjhaveri/indicNLP) package provides a collection of utilities to process Gujarati text: sentence splitting, word tokenization, stopword detection, stemming, part-of-speech tagging, word-variation identification (text normalization), named-entity recognition, and document classification. `NLP for Gujarati` (https://github.com/goru001/nlp-for-gujarati) provides word-tokenization and a pre-trainded language model. A wordnet (https://github.com/sagarpanchal8793/Gujarati-Wordnet) and a named-entity recognition tool (https://github.com/nikitsaraf/Named-Entity-Tagger-Gujarati) are also available. A repository with Gujarati stopwords can be downloaded from https://github.com/gujarati-ir/Gujarati-Stop-Words. The tool `ReadabilityScore` provides an indication of the readibility of a text written in Gujarati (https://github.com/Somsubhra/ReadabilityScore).

Bilingual resources: A crowd-sourced Gujarati–English dictionary (http://www.seas.upenn.edu/~nlp/resources/TACL-data-release/dictionaries.tar.gz) has been made available by Pavlick et al. (2014).

The *Glosbe* bilingual concordancer can be used online from the website: https://glosbe.com/en/gu.

Some machine translation systems are freely available on the Internet:

- Yandex translate: https://translate.yandex.com
- Google Translate: https://translate.google.com/#gu/en/

2.5.5 Challenges for corpus-based MT from English

In addition to the scarcity of Gujarati–English text, the main challenges probably come from generating Gujarati output with the right word order, namely at two levels:

- General sentence level: While English is a subject–verb–object language, Gujarati is a subject–object–verb language, which may lead to long-range reordering during translation.
- Phrase level: the use of post-positions vs. pre-positions, or pre-modifying relative clauses in Gujarati will require to re-order morphemes when translating English words.

This two-level re-ordering may be challenging for MT systems. Neural MT may be better at dealing with long-range reordering than statistical MT, though. Another challenging aspect may be to determine the correct case in Gujarati, which is not marked in English, and which affects morphology.

Another problem may be that the way in which verbs agree with their arguments (agent, patient, etc.) depends on verbal aspect, which is sometimes hard to determine in English text.

Generating the right word order in a question could be also challenging, given that Gujarati does not place interrogative phrases initially as English does, but instead it places them where the corresponding phrase in the affirmative sentence would be, and the order of words in polar questions does not change with respect to affirmative sentences.

2.6 Hausa (ha, hau)

2.6.1 Factsheet

Hausa is an Afro-Asiatic language, as Tigrinya (§ 2.14) in the Chadic family and Afaan Oromoo (§ 2.1) in the Cushitic family. According to Wikipedia it is spoken by about 44 million people as a first language, and perhaps by about 20 million people as a second language (as it has reached the status of *lingua franca* in West Africa.[10] It is official in Nigeria, together with Igbo (§ 2.7) and Yoruba (§ 2.16), in Niger and in Ghana.

Hausa is written mostly with a modified Latin alphabet. There is a system called Boko, which contains a number of special characters (ɓ,Ɓ,ɗ,Ɗ,ƙ,Ƙ, etc.), which is however not used by broadcasters such as the BBC. Hausa is a tonal language, with three tones, but most writing does not mark tones (when it does, accents are used).

2.6.2 Contrasts with English

Examples are written in the simplest possible transcription, disregarding tone marks and avoiding special Latin characters. A good part of the examples are taken from Jaggar (2001) and some from Campbell and King (2010). This is not an exhaustive list of contrasts: not all contrasts in the *World Atlas of Linguistic Structures* are documented.

[10] Campbell and King (2010) report 25 and 12 million respectively.

Morphology			
Feature	**Value in English**	**Value in Hausa**	**Examples**
Verb inflection	Verbs may inflect for person and tense.	Verbs consist of two parts: the person-aspect complex and the the main verb itself. The verb system is aspect-based rather than tense-based.	*Audu ya fita* 'Audu went out', lit. 'Audu *3rd-masculine-perfective* go-out'; *Audu yana fita* 'Audu is/was going out', lit. 'Audu *3rd-masculine-imperfective* go-out'.
Nominal morphology (nouns and adjectives)	Mainly suffixing where present	Complex system, including suffixing and partial reduplication.	*giwa* 'elephant', pl. *giwaye*; *gari* 'town', pl. *garuruwa*; *makaranta* 'school', pl. *makarantu*.
Number of genders	Three, only in third-person singular pronouns and possessives	Two, assigned semantically but also formally	*sabon iyali* 'new family', *sabuwar duniya* 'new world'
Demonstratives	Two sets	Four sets (but only two sets unless tones are marked)	*wannan riga* 'This gown [near me]' or 'That gown [near you]' *wancan riga* 'That gown [far from me and you]' or 'That gown [very far from me and you]'

Syntax			
Feature	**Value in English**	**Value in Hausa**	**Examples**
Order of Genitive and Noun	No dominant order	Noun-Genitive	*abooki-n ubaa* 'The friend of the father' (where *abooki* is masc.), approx. 'friend-his father'; *goona-r ubaa* 'The field of the father', approx. 'Field-his father' (where *goona* is feminine); note the use of marking on the *modified* noun, rather than on the modifying genitive.[11].
Definite article	Word different from demonstrative	Bound suffix	*mota* 'car', *motar* 'the car'; *yaro* 'boy', *yaron* 'the boy'
Pronominal subjects	Obligatory	Not obligatory	*Yana koyon Hausa* 'He is/was learning Hausa', lit. '*3rd-masculine-imperfective* learning Hausa', no *He* pronoun.
Order of Demonstrative and Noun	Demonstrative–noun	Mixed	*yaron-nan* 'This boy [we just talked about]' but *wannan yaro* 'This boy [general]'
Order of numeral and noun	Numeral–noun	Noun–numeral	*watanni uku* 'Three months', lit. 'Months three'
Possessive affixes	No possessive affixes	Possessive suffixes, extensively used in genitive constructs	*kudi-n Audu* 'Audu's money', lit. 'Money-his Audu', *baba-r yarinyar* 'The girl's mother', lit. 'Mother-her girl'
Verbal and nominal coordination	Same	Different	*Shayi* ***da*** *sukari* 'Tea **and** sugar', 'Tea **with** sugar'; *Ya zo* ***kuma*** *ta gan shi* 'He came **and** she saw him'
Verbal clause negation	Subject–Negation–Verb–Object (generally)	Subject–Negation–Verb–Object–Negation (generally)	*Yara za su karanta littafin.* 'The boys will read the book', *Yara* ***ba*** *za su karanta littafin* ***ba****.* 'The boys will not read the book'.
Position of Interrogative Phrases in Content Questions	Initial interrogative phrase	Mixed: initial or non-initial	Intital in ***Yaushe*** *za ka dawo?* '**When** will you come back?' but also in situ in *Ya tafi* ***yaushe****?* '**When** did he leave' lit. 'He left **when**?'

[11]Campbell and King (2010) call this the *compound* state of the noun. It is used in many other grammatical constructs.

corpus	doc's	sent's	en tokens	ha tokens
Tanzil (Quran) v1	15	0.1M	2.8M	3.0M
GNOME v1	62	19.0k	0.2M	0.1M
KDE4 v2	3	1.5k	14.3k	4.9k
Ubuntu v14.10	7	0.3k	9.2k	0.8k
total	87	120.8k	3.0M	3.1M

Table 6: Distribution of the sentences in the Opus Hausa–English corpus.

2.6.3 Corpora

Monolingual corpora: The monolingual dump of the Hausa wikipedia is periodically made available at https://dumps.wikimedia.org/hawiki/. The A5 Hausa Umarnin Uwa corpus (https://corpora.uni-hamburg.de/hzsk/de/islandora/object/spoken-corpus:a5hausaumarninuwa) is extracted from the Umarnin Uwa film transcripts and contains 47 transcripts with a total of 10,194 tokens in Hausa. It provides information including automatic part-of-speech tagging, speaker and extralinguistic information, foreign words and code-switching. Similarly, the A5 Hausa News corpus (http://hdl.handle.net/11022/0000-0000-82AC-B) is extracted from a collection of news articles from the online news service of DW and contains 4 texts with a total of 2,017 tokens. A Hausa conversation transcriptions corpus is provided by CoCooN (https://cocoon.huma-num.fr/exist/crdo/search2.xql?subject=http%3A%2F%2Flexvo.org%2Fid%2Fiso639-3%2Fhau).

Bilingual corpora: Four corpora available at OPUS, and three of them come from free/open-source software projects (details are available in table 6):

- Tanzil (Quran): http://opus.nlpl.eu/download.php?f=Tanzil/v1/moses/en-ha.txt.zip
- GNOME: http://opus.nlpl.eu/download.php?f=GNOME/v1/moses/en-ha.txt.zip
- Ubuntu: http://opus.nlpl.eu/download.php?f=Ubuntu/v14.10/moses/en-ha.txt.zip
- KDE: http://opus.nlpl.eu/download.php?f=KDE4/v2/moses/en-ha.txt.zip

2.6.4 Resources

Monolingual resources: A collection of stopwords is provided by the `more-stoplists` project (https://github.com/dohliam/more-stoplists) for African languages such as Swahili, Yoruba, Hausa or Zulu.

Bilingual resources: The Glosbe bilingual concordancer (https://glosbe.com/ha/en) offers translations into English for Hausa words in context.

To the best of our knowledge, the only machine translation system available online is Google Translate: https://translate.google.com/#ha/en/

2.6.5 Challenges for corpus-based MT to English

In addition to the scarcity of Hausa–English text, the main translation challenges are:

- The fact that the verbal system in Hausa is dominated by aspect (perfective or imperfective) rather than by tense means that inferring the correct English tense will be difficult.
- Complexity in nominal morphology, where both suffixing and partial reduplication are used for plurals.
- The possibility of not specifying the pronominal subject may be a source of ambiguity in some cases.
- The absence of tone markings may make some translations into English difficult. For instance, *wannan* could be 'this' (*wannàn*) or 'that' (*wànnan*).
- Lack of adherence to a single standardized orthography (for instance, as regards the use of special Boko characters such as ɓ, ɗ, etc.).

2.7 Igbo (ig, ibo)

2.7.1 Factsheet

According to Wikipedia, Igbo is the principal native language of the Igbo people, an ethnic group of southeastern Nigeria. The language has approximately 44 million speakers. Igbo is official in Nigeria, where Hausa (§ 2.6) and Yoruba (§ 2.16) are also large languages. It belongs to a very large family, the Niger–Congo family, as Swahili (§ 2.13), and more specifically to the Volta–Niger group, as Yoruba (§ 2.16).

Igbo is written with the Latin script, mostly in the *Ọnwụ* script, which contains the modified letters *ị*, *ṅ*, *ọ*, *ụ*. Igbo is a tone language. Tones are not consistently indicated, despite their importance in distinguishing words (for instance, *oke* 'male' / *okè* 'limit' / *òkè* 'portion') (Ugochukwu, 2004). If they are, acute and grave accents are used. Reputed news outlets such as the BBC Igbo news (https://www.bbc.com/igbo) do not always write Igbo tones, similarly to what happens with Yoruba, see page 77.

2.7.2 Contrasts with English

Some examples are taken from Ugochukwu (2004) or from Glosbe (https://glosbe.com/en/ig/).

Morphology			
Feature	**Value in English**	**Value in Igbo**	**Examples**
Number of genders	Three (sex, based, only in singular pronouns and possessives)	None	English has *she*, *he*, and *it*, where Igbo uses only *ọ*.
Coding of Nominal Plurality:	plural suffix in most cases	no morphological plural	Plural expressed by numerals, by context, or by reduplication (non specific): *ụlo àtọ* 'Three houses', lit. 'House three' ; *onono onono* 'bottles', 'many bottles', lit. 'bottle bottle' (Anagbogu, 1995)

Verbs			
Feature	**Value in English**	**Value in Igbo**	**Examples**
Verbal person and number marking	Marked together (when marked)	Neither is marked	*Ọ nà-azà ụlọ̀* 'He is sweeping the house'; *Anyị nà-azà ụlọ̀* 'We are sweeping the house'.
Morphological imperative	No distinction between 2nd person singular and plural	Different forms	*Tàa àchịchà!* ('Eat biscuits!', singular), *Tàanụ̀ àchịchà!* ('Eat biscuits', plural).

Word order and syntax			
Feature	**Value in English**	**Value in Igbo**	**Examples**
Order of genitive and noun	No dominant order	Noun-genitive	*ụlọ Nna* 'The Father's house' (lit. 'house Father'); *ụlọ ya* 'his house' (lit. 'house him').
Dominant order of adjective and noun	Adjective–noun	Noun–adjective	*àgbà ọhụu* 'New Testament', lit. 'Testament New'; *àgwà ọcha* 'white bean', lit. 'bean white'
Order of demonstrative and noun	Demonstrative–noun	Noun–demonstrative	*Nwoke ahụ* 'That man', lit. 'Man that';
Order of numeral and noun	Numeral–noun	Noun–numeral	*ụlo àtọ* 'Three houses', lit. 'House three'
Position of interrogative phrases in content questions	Initial interrogative phrase	Not initial interrogative phrase	*Aha gị bụ gini?* 'What is your name', lit. 'Name you is what?', same order as *Aha gị bụ John* 'Your name is John'
Passive constructions	Present	Absent	Igbo has an impersonal pronoun *a/e* instead: *E riri yaa* 'Yam was eaten', lit. 'Someone ate yam', parallel to *Anyị riri yaa* 'We ate yam' (Akinremi, 2013).
Yes/no interrogatives	Different order from affirmative	Interrogative intonation only	*Ị na-eri nri?* 'Do you eat bread'; compare *Ị na-eri nri* 'You eat bread'
Nominal and Locational Predication	Identical	Different	*John bụ nwoke* 'John is a man', but *John nọ n'ụlọ* 'John is at home'.

2.7.3 Corpora

Monolingual corpora: The monolingual dump of the Igbo wikipedia is periodically made available at https://dumps.wikimedia.org/igwiki/. The *igTenTen* is available from the sketch platform, but requires a subscription: https://www.sketchengine.eu/igtenten-igbo-corpus/.

Bilingual corpora: In addition to the BBC (https://www.bbc.com/igbo, we have only found TWR360 to produce content in Igbo (https://www.twr360.org/, although mostly multimedia content). These outlets may be interesting sources from which to obtain monolingual corpora.

Two corpora are available at OPUS, from two different free/open-source software projects (details are available in table 7):

- GNOME: http://opus.nlpl.eu/download.php?f=GNOME/v1/moses/en-ig.txt.zip

corpus	doc's	sent's	en tokens	ig tokens
GNOME v1	72	23.8k	200k	300k
Ubuntu v14.10	8	0.6k	10.2k	4.6k
total	80	24.4k	210.2k	304.9k

Table 7: Distribution of the sentences in the OPUS Igbo–English corpus.

- Ubuntu: http://opus.nlpl.eu/download.php?f=Ubuntu/v14.10/moses/en-ig.txt.zip

2.7.4 Resources

Monolingual resources: No monolingual resources have been found for the Igbo language.

Bilingual resources: Some machine translation systems are freely available on the Internet:

- xLingua: http://igbo.xlingua.net/ig/
- Google Translate: https://translate.google.com/#ig/en/

Some collections of bilingual phrases are available at:

- Igbo English.com: http://www.igboenglish.com/
- Ilanguages.org: http://ilanguages.org/igbo_phrases.php

The *Glosbe* bilingual concordancer can be used online from the website: https://glosbe.com/en/ig

2.7.5 Challenges for corpus-based MT from English

One of the main challenges when translating English into Igbo is the lack of parallel corpora. Apart from this, Igbo has a simpler morphology than English. One of the most relevant challenges as regards morphology is the translation of plural nouns into Igbo: on the one hand, plural is not always marked morphologically, and, on the other hand, there are two different strategies to mark plurals, sometimes even through reduplication.

Word order is locally different between both languages. However, most of the differences happen in a small context in which most corpus-based MT systems should be able to succeed, for example, noun–genitive order or noun–adjective order (see section 2.7.2).

One of the challenges to be faced by MT systems is to determine the initial position of an interrogative sentence in Igbo, given the fact that, in this language, interrogative phrases are placed where the corresponding phrases would be in an affirmative sentence.

Another relevant challenge may be to adequately translate English passive sentences, as passives do not exist in Igbo.

2.8 Korean (ko, kor)

2.8.1 Factsheet

Korean —usually considered a language isolate despite having relatives in the Koreanic language family such as the Jeju language spoken in the Jeju province of South Korea— has more than 70 million speakers, according to Wikipedia, most of them in North and South Korea, where it is official. It is also official in the Yanbian prefecture of China.

It is written in Hangul,[12] a five-century-old syllabary unique to Korean[13] which has however undergone extensive reform since it was adopted in the 16th century.

Decades of political separation between North and South Korea have resulted in two different standards for the Korean language. As a result, there are some small graphical differences in the representation of some sounds, some differences in inflected forms of words, lexical differences, and even spacing differences, which may affect tokenization.

Project GoURMET is mostly interested in the North Korean variety. General Korean–English corpora, in addition to lacking examples of Northern grammar, spelling, and spacing, may not contain Northern named-entities or propose Southern variations for their English counterparts.

2.8.2 Contrasts with English

Korean is quite typical of a subject–object–verb language: it has postpositions (some of which are also called *particles*) instead of English prepositions, modifiers precede the modified clauses, etc.

Examples are given in transliteration.

[12] Also called Chosongul in North Korea

[13] Almost unique: it is also used to write Ciacia, a language spoken around the city of Baubau in Indonesia, and also for the related Jeju language.

Syntax			
Feature	**Value in English**	**Value in Korean**	**Examples**
Order of subject, object and verb	Subject–verb–object	Subject–object–verb	*Naneun ramyeoneul meogeotda* 'I ate ramen' [I-TOPIC ramen-OBJECT ate]
Order of Adposition and Noun Phrase	Prepositions	Postpositions and case suffixes	*jib-**eseo*** 'From the house'; *jib-**e*** 'To the house'.
Order of Genitive and Noun	No dominant order	Genitive–Noun	*namja[-ui] jadongchayeyo* 'The man's car', lit. 'Man[GENITIVE] car'
Order of Relative Clause and Noun	Noun–Relative clause	Relative clause–Noun	*Naega sassdeon cha* 'The car that I bought', lit. 'I bought-REL car'
Evidentiality	No morphological evidentials	Morphological evidentials (direct/indirect)	Contrast: *Peterga Maryga jandago **malhayeossda*** 'Peter said that Mary was sleeping' (direct evidence); *Peterga Maryga jandago **malhadeora*** (indirect evidence coded in the verb 'said': 'I heard that Peter said…') Song (2010)
Interrogative phrases in content questions	Initial	Not initial	*jon-eun **leondeon-e** salgo issseubnida* 'John lives **in London**'; *jon-eun **eodieseo** salgo issseubnikka?* '**Where** does John live?'

Function words			
Feature	**Value in English**	**Value in Korean**	**Examples**
Definite article	Different from demonstrative	No definite article	*jadongcha* 'car' or 'the car'
Indefinite article	Different from 'one'	No indefinite article	*jadongcha* 'car' or 'a car'
Distance Contrasts in Demonstratives	Two-way contrast (near speaker, rest)	Three-way contrast (near speaker, near hearer, far from both)	
Comitatives and instrumentals	Identical	Different	*-[eu]ro* is the instrumental 'with' (*peneuro* 'with a pen'); *-[g]wa* is the comitative 'with' (*chinguwa* 'with a friend').

2.8.3 Corpora

Monolingual corpora: The KAIST corpus (http://semanticweb.kaist.ac.kr/home/index.php/KAIST_Corpus) collects a series of sub-corpora including monolingual and bilingual data. Some of these sub-corpora are domain-specific. In addition, some corpora are augmented with morphologic and syntactic annotations.

A number of corpora obtained from transcriptions is available; namely:

- Korean Telephone Conversations Transcripts corpus (https://catalog.ldc.upenn.edu/LDC2003T08): consists of 100 telephone conversations in Korean transcribed;
- the Korean Broadcast News Transcripts corpus (https://catalog.ldc.upenn.edu/LDC2006T14): collects 18 text files containing transcripts from Voice of America satellite radio news broadcasts in Korean;
- the Ryu Spoken Corpus (https://childes.talkbank.org/access/EastAsian/Korean/Ryu.html) and the Jiwon Spoken Corpus (https://childes.talkbank.org/access/EastAsian/Korean/Jiwon.html): transcriptions of children speaking in Korean.

Korean Newswire corpus (https://catalog.ldc.upenn.edu/LDC2000T45) and Korean Newswire corpus second edition (https://catalog.ldc.upenn.edu/LDC2010T19) are a collection of Korean Press Agency news articles. The Korean Treebank Annotations Version 2.0 (https://catalog.ldc.upenn.edu/LDC2006T09) is an electronic corpus of Korean texts annotated with morphological and syntactic information; original texts come for the Korean Treebank 2.0 and were selected from The Korean Newswire corpus. The C4Corpus (https://dkpro.github.io/dkpro-c4corpus/) is extracted from the on-line available CommonCrawl, a massive crawl of documents from the Internet. The monolingual dump of the Korean wikipedia is periodically made available at https://dumps.wikimedia.org/kowiki/. Sejong corpus (http://universal.elra.info/product_info.php?cPath=42_43&products_id=1975).

Some corpora are only available after paying a fee. The Qualified POS Tagged Corpus (http://catalog.elra.info/en-us/repository/browse/ELRA-W0034/) is produced by KAIST KORTERM, containing 1,020,000 eojeols (Korean terms). A Korean lexicon (http://catalog.elra.info/en-us/repository/browse/ELRA-L0044/) is also available consisting of 31,476 compound nouns in Korean.

In addition to BBC (https://www.bbc.com/korean) the following international media outlets produce content in Korean: Global Voices (https://ko.globalvoices.org/), The Voice of America (https://www.voakorea.com/) , China Plus (formerly China Radio International, http://korean.cri.cn/), Vatican Radio (https://www.vaticannews.va/ko.html), NHK World (https://www3.nhk.or.jp/nhkworld/ko/), and TWR360 (https://www.twr360.org/, although mostly multimedia content). These outlets may be interesting sources from which to obtain monolingual corpora.

Bilingual corpora: The JHE Korean–English evaluation data (https://zenodo.org/record/891295#.XKxazkPgpD8) is a small parallel corpus for machine translation evaluation by Park et al. (2016). The Korean English News v1 (https://github.com/jungyeul/korean-parallel-corpora/tree/master/korean-english-news-v1) is a parallel corpus from news articles.

Eight corpora are available at OPUS, from three free/open-source software projects (details are available in table 8):

corpus	doc's	sent's	en tokens	ko tokens
OpenSubtitles v2018	1774	1.6M	12.4M	7.8M
GNOME v1	1581	0.6M	3.1M	2.4M
OpenSubtitles v2016	458	0.4M	3.3M	2.1M
Tanzil (Quran) v1	15	93.6k	2.8M	1.6M
Tatoeba v2	1	0.9k	3.6M	9.0k
KDE4 v2	597	87.3k	0.6M	0.4M
Ubuntu v14.10	357	91.9k	0.6M	0.3M
PHP v1	3220	50.7k	0.5M	0
OpenSubtitles v2013	28	28.1k	0.2M	0.2M
GlobalVoices v2017q3	348	8.2k	0.2M	0.2M
OpenSubtitles v2012	17	16.4k	0.1M	99.8k
OpenSubtitles v2011	8	6.5k	58.8k	38.9k
total	8404	3.0M	27.1M	15.1M

Table 8: Distribution of the sentences in the Opus Korean–English corpus.

- Open Subtitles: http://opus.nlpl.eu/download.php?f=OpenSubtitles/v2018/moses/en-ko.txt.zip
- GNOME: http://opus.nlpl.eu/download.php?f=GNOME/v1/moses/en-ko.txt.zip
- Tanzil (Quran): http://opus.nlpl.eu/download.php?f=Tanzil/v1/moses/en-ko.txt.zip
- Tatoeba: http://opus.nlpl.eu/download.php?f=Tatoeba/v2/moses/en-ko.txt.zip
- KDE4: http://opus.nlpl.eu/download.php?f=KDE4/v2/moses/en-ko.txt.zip
- PHP: http://opus.nlpl.eu/download.php?f=PHP/v1/moses/en-ko.txt.zip
- GlobalVoices: http://opus.nlpl.eu/download.php?f=GlobalVoices/v2017q3/moses/en-ko.txt.zip
- Ubuntu: http://opus.nlpl.eu/download.php?f=Ubuntu/v14.10/moses/en-ko.txt.zip

The *KAIST Corpus*[14] consists of a collection of corpora in Korean, among which we can find several parallel corpora:

- Corpus7: http://semanticweb.kaist.ac.kr/home/index.php/Corpus7
- Corpus9: http://semanticweb.kaist.ac.kr/home/index.php/Corpus9
- Corpus10: http://semanticweb.kaist.ac.kr/home/index.php/Corpus10
- Newspaper corpus: http://semanticweb.kaist.ac.kr/download/form.php?cid=12

By subscribing to the Sketch Engine it is possible to access two more corpora:

[14] http://semanticweb.kaist.ac.kr/home/index.php/KAIST_Corpus

- Timestamped JSI web corpus (https://www.sketchengine.eu/jozef-stefan-institute-newsfeed-corpus/#toggle-id-1)
- koTenTen: Corpus of the Korean Web (https://www.sketchengine.eu/kotenten-korean-corpus/)

Some additional corpora are only available after paying a fee, such as the Collins Multilingual database (MLD) - PhraseBank (http://catalog.elra.info/en-us/repository/browse/ELRA-T0377/) and WordBank (http://catalog.elra.info/en-us/repository/browse/ELRA-T0376/). The Computer Science Database (http://catalog.elra.info/en-us/repository/browse/ELRA-T0366/) is a 76,272 entries in Korean and in English in the field of computer science. A Multilingual Corpus (http://catalog.elra.info/en-us/repository/browse/ELRA-W0035/) of expressions in Korean is available, containing the equivalents in Chinese and English. A Biology Database (http://catalog.elra.info/en-us/repository/browse/ELRA-T0365/) is also available, consisting of 31,884 entries in Korean and English in the field of biology.One of the

2.8.4 Resources

Monolingual resources: Some monolingual resources are provided within the Sketch Engine (https://www.sketchengine.eu/user-guide/user-manual/corpora/by-language/korean-text-corpora/); it is mandatory to be a subcriber to access them: Korean Word Sketch, Korean thesaurus, Korean word lists, Korean concordance, and N-grams in Korean.

As regards language technologies, there is a number of natural-language-processing tools freely available for Korean. `KoNLPy` (https://github.com/konlpy/konlpy) is a tools that enables morphological analysis and part-of-speech tagging. In addition, it provides a collection of corpora and dictionaries. The library `open-korean-text` (https://github.com/open-korean-text/open-korean-text) allows text processing with Java. Namely, it provides text normalization and tokenization. `RKMA` (https://github.com/youhyunjo/rkma) is an R library for morphological analysis. `Kormoran` (https://github.com/shineware/komoran-2.0) is one of the most popular morphological analysers for Korean. There are wrappers for several programming languages: Java, Python, R, etc.

Bilingual resources: Some machine translation systems are freely available on the Internet:

- Yandex translate: https://translate.yandex.com
- Bing translate: https://www.bing.com/translator
- Google Translate: https://translate.google.com/#ko/en/

A crowd-sourced Korean–English dictionary (http://www.seas.upenn.edu/~nlp/resources/TACL-data-release/dictionaries.tar.gz) has been made available by Pavlick et al. (2014). In addition to this, the Korean Propbank (https://catalog.ldc.upenn.edu/LDC2006T03), provides a semantic annotation of the Korean–English Treebank Annotations and Korean Treebank Version 2.0.

2.8.5 Challenges for corpus-based MT to English

One of the main challenges when translating from Korean into English is the difference in the structure of these languages. While English is a subject–verb– object language, Korean is a subject–object–verb language; this may require long-range reordering during translation.

Moreover, there are linguistic phenomena that exist in Korean and do not have an equivalent in English, which may make difficult to decide which is the best choice when translating into English. Namely:

- the fact that Korean uses evidentiality marks, a phenomenon that does not exist in English and may not be easy to translate;
- the lack of articles (both definite and indefinite) in Korean, which makes difficult to choose the right article in English;
- the difference in the definition of the distance levels between speakers.

Finally, the lack of corpora and resources covering specifically the Northern variety of Korean, will surely make it more challenging to deal with the specific grammar, spelling, spacing, and named entities found in North Korean.

2.9 Kurdish (ku, kur, kmr, ckb, sdh)

2.9.1 Factsheet

According to Wikipedia, Kurdish is a continuum of languages spoken by the Kurds in Western Asia. These languages belong to the Indo-Iranian family inside the Indo-European family, as do two other languages in this project, namely Gujarati (§ 2.5) and Punjabi (§ 2.11).

Kurdish languages form three groups known as Northern Kurdish (Kurmanji, individual code `kmr`), Central Kurdish (Sorani, individual code `ckb`), and Southern Kurdish (Palewani or Kirmashani, individual code `sdh`); they are not mutually intelligible without learning. Studies as of 2009 estimate between 8 and 20 million native Kurdish speakers in Turkey.

The majority of Kurds speak Northern Kurdish (Kurmanji) in Turkey, Syria, northern Iraq, and northwest and northeast Iran. Central Kurdish is spoken by an estimated 7 million Kurds in the Iraqi Kurdistan and the Iranian Kurdistan Province. Kurmanji is written in the Latin script, whereas Sorani is mainly written in a modified version of the Arabic–Persian script, although the Latin script is often used in particular contexts such as messaging applications.

The *endonymic glossonym* (name of the language as given by its own native speakers) for Central Kurdish is *Kurdîy nawendî* and also *Soranî*. The endonymic glossonym for Northern Kurdish is *Kurmancî* (Kurmanji).

Two related languages are Zaza–Gorani and Persian. The last one is better resourced than any Kurdish variety.

The Internet geographic top-level domain for Kurdistan Region of Iraq is `.krd`.

2.9.2 Contrasts with English

The following description may be more accurate for Sorani, as this is the Kurdish language for which the World Atlas of Linguistic Structures reports more contrasts with English. Examples —where given— are taken from Wikipedia and from Thackston (2006a) and Thackston (2006b).

Morphology			
Feature	**Value in English**	**Value in Kurdish**	**Examples**
Morphological cases	Two, only pronouns and some closed-class words	Thackston (2006a) reports four (nominative, oblique, construct,[15] and vocative) for for Kurmanji, but Thackston (2006b) does not explicitly discuss case, except for ergativity.	
Gender	Two, only 3rd-person singular pronouns	Thackston (2006a) reports two genders for Kurmanji, but Thackston (2006b) does not report gender for Sorani.	
Morphological indicators in verb	Tense, person, number	Verbs inflect with person, number, mood, tense, polarity. *bíchim* (Kurmanji) / *ez bíçim* (Sorani) 'That I go'	
The transitive verb agrees with…	…the A argument ('agent')	…the A argument in imperfective verb forms, and the P argument ('patient') in perfective verb forms (split ergativity)	Kurmanji, perfective Thackston (2006a): *wî ez dîti**m*** 'he saw me' (lit. 'he me saw-me') vs. *wî em dîti**n*** 'he saw us' (lit. 'he us saw-us')

[15]The construct or *ezafe* is not properly a case.

Function words			
Feature	**Value in English**	**Value in Kurdish**	**Examples**
Definite Articles	Definite word distinct from demonstrative	Definite affix (Sorani, not Kurmanji)	*pyâw* 'man', *pyâwaká* 'the man'
Indefinite Articles	Indefinite word distinct from 'one'	Indefinite affix	*miróvek* 'a man' (*mirov* '(the) man')) (Kurmanji, Thackston (2006a)); *pyâwèk* 'a man' (*pyâw* 'man') (Sorani, Thackston (2006b)).

Word order			
Feature	**Value in English**	**Value in Kurdish**	**Examples**
Order of Subject, Object and Verb	Subject–Verb–Object	Subject–Object–Verb	*Ew wî mirovî dibîne* 'He sees that man', lit. 'He that man sees' (Kurmanji, Thackston (2006a))
Order of Genitive and Noun	No dominant order	Noun–Genitive	*jega-y pasa* 'The king's place', lit. 'place-CONSTRUCT king' [As many other languages, even so unrelated as Hausa (2.6) do, Kurdish marks the *modified* noun with the ending *-y*, sometimes called the *construct* state or *ezafe*.]
Order of Adjective and Noun	Adjective–Noun	Noun–Adjective	*pyaw-ı čak* [Note that, as in genitive constructs, the modified noun is in the construct state (carries an *ezafe* ending).]

Other syntax			
Feature	**Value in English**	**Value in Kurdish**	**Examples**
Negative	Negative particle	Negative affix in verb (except in copulas)	Sorani: *dáchim* 'I go', *nádachim* 'I don't go'; Kurmanji: *ez védikim* 'I do', *ez venákim* 'I don't do'.

2.9.3 Corpora

Monolingual corpora: The Kurdish-BLARK project (https://github.com/hosseinhassani/Kurdish-BLARK) contains monolingual corpora of Kurmanji (around 12,000 words) and Sorani (around 273,000 words) released under a GNU AGPL v3 license.

In 2019, the AsoSoft Sorani corpus by Veisi et al. (2019) has been released (https://github.com/AsoSoft/AsoSoft-Text-Corpus) under non-comercial license. It contains 75 million tokens.

The Voice of America has two services for Kurdistan (https://www.dengiamerika.com/, Arabic script, probably the Sorani variety, and https://www.dengeamerika.com/, Latin script, probably

the Kurmanji script.). TWR360 (https://www.twr360.org/) has a Sorani site, mostly with multimedia content. Global Voices has what looks like a Sorani site which it calls Kurdish (https://ku.globalvoices.org/.

There are two independent online newspapers in Iraqi Kurdistan written in Sorani: *Awene* (http://www.awene.com/) and *Hawlati* (http://www.hawlati.co/).

The Sorani Wikipedia is available on https://ckb.wikipedia.org/wiki/. As of April 2019 it contains 22,894 entries (see stats on https://stats.wikimedia.org/EN/SummaryCKB.htm). A monolingual dump is periodically made available at https://dumps.wikimedia.org/ckbwiki/.

The Kurmanji Wikipedia is available on https://ku.wikipedia.org/. As of April 2019 it contains 24,404 entries (see stats on https://stats.wikimedia.org/EN/SummaryKU.htm). A monolingual dump is periodically made available at https://dumps.wikimedia.org/kuwiki/.

The Iranian Studies site at Harvard University contains selected readings in Sorani Kurdish (https://sites.fas.harvard.edu/~iranian/Sorani) and Kurmanji Kurdish (https://sites.fas.harvard.edu/~iranian/Kurmanji). They are distributed under a CC BY-NC-ND 4.0 license. The site also contains reference grammars in English for both languages (Thackston, 2006a,b).

Bilingual corpora: Bianet (Ataman, 2018) contains 6,486 English–Kurmanji parallel sentences in the news domain. It can be downloaded from https://d-ataman.github.io/bianet and it is distributed under a CC-BY-SA-4.0 license. It also contains parallel sentences for English–Turkish and Turkish–Kurmanji.

Opus (http://opus.nlpl.eu) has a corpus Kurmanji–English of approximately 300,000 sentences of which around 93,000 are Quran translations and the remaining sentences belong to the documentation of Ubuntu, GNOME and KDE.

Kurdish News Network (https://www.knnc.net/) contains news articles in Sorani and English but it needs to be checked whether they are actually parallel texts. A similar study needs to be carried out for the Kurmanji and English articles published by the Hawar News Agency (https://www.hawarnews.com/kr/).

Another Iraqi media network, Rudaw Media Network, has a newspaper (http://www.rudaw.net) published in English, Kurmanji and Sorani, as well as radio and TV channels.

Google Translate supports Kurmanji–English translation since 2016; Sorani is not supported as of April 2019 (but see http://www.rudaw.net/english/kurdistan/110220193). Inkurdish (https://www.inkurdish.com/) is another commercial system. Apertium (see below) also has Kurmanji–English translation. In 2016, Translators Without Borders released Apertium-based machine translation systems for Kurmanji–English and Sorani–English (see https://translatorswithoutborders.org/translators-without-borders-develops-worlds-first-crisis-specific-machine-translation-system-kurdish-languages/; the corresponding resources are free/open source: https://github.com/apertium/apertium-kmr-eng, https://github.com/apertium/apertium-ckb-eng).

2.9.4 Resources

Monolingual resources: The Kurdish-BLARK project (https://github.com/hosseinhassani/Kurdish-BLARK) contains some tools (published under a GNU AGPL v3 license) such as a transliterator from Persian/Arabic texts into Latin script, a tokenizer, a stemmer to find Kurmanji and

Sorani stems, a word-level translator from Kurmanji to Sorani (and vice versa) based on a bilingual dictionary, an a Kurdish proper names recognizer. The tools have been described by Hassani (2018).

KurLex (https://gforge.inria.fr/scm/viewvc.php/alexina/kurlex/trunk/) is a morphological lexicon for Kurmanji Kurdish as described by Walther et al. (2010). KurLex is distributed under the LGPL-LR license; the release 0.0.1 can be downloaded from https://gforge.inria.fr/frs/?group_id=482.

SoraLex (https://gforge.inria.fr/scm/viewvc.php/alexina/soralex/trunk/) is a morphological lexicon for Sorani Kurdish as described by Walther and Sagot (2010). SoraLex is distributed under the LGPL-LR license; the release 0.0.1 can be downloaded from https://gforge.inria.fr/frs/?group_id=482.

An Crúdabán (http://crubadan.org/writingsystems) has lists of words and bigrams for different variants and scripts of Kurdish; files containing a small number of URLs that were used to compile the words are also included.

Apertium includes linguistic data for Kurmanji (https://github.com/apertium/apertium-kmr) and Sorani (https://github.com/apertium/apertium-ckb): morphological analysers and part-of-speech taggers. See 'Bilingual Resources' below for more detail.

The paper by Esmaili (2012) and the webpage "Building a Kurdish Language Corpus: an overview of the Technical Problems" (http://ggautierk.free.fr/e/icem_98.htm) by Gérard Gautier discuss some interesting points about linguistic resource development for Kurdish.

The first syntactically annotated corpus of Kurmanji Kurdish contains approximately 10,000 words and was developed by Gökırmak and Tyers (2017) and released under a Creative Commons License Attribution-ShareAlike 4.0 International at https://github.com/UniversalDependencies/UD_Kurmanji-MG.

Bilingual resources: Apertium also contains bilingual data for Kurmanji–English (https://github.com/apertium/apertium-kmr-eng) and Sorani–English (https://github.com/apertium/apertium-ckb-eng): bilingual dictionaries and structural transfer-rules. The final report (see http://wiki.apertium.org/wiki/Kurmanji_and_English/Final_report) by the Google Summer of Code 2016 student that created most of the data stated that the Kurmanji–English system contained around 17,000 dictionary entries, 157 paradigms and 23 transfer rules. All data are published under a GPL3 license.

The Iranian Studies site at Harvard University contains a Sorani Kurdish vocabulary in English (https://sites.fas.harvard.edu/~iranian/Sorani) and also a Kurmanji Kurdish vocabulary in English (https://sites.fas.harvard.edu/~iranian/Kurmanji). They are distributed under a CC BY-NC-ND 4.0 license.

2.9.5 Challenges for corpus-based MT to English

The main challenges when translating from both Sorani and Kurmanji Kurdish to English come from grammatical differences:

- agglutination in noun-based phrases to represent gender and case (Kurmanji) or pronominal suffixes (Sorani);

- radically different sentence and phrase structures —position of object, obliques, and verb; use of postpositional elements and circumpositions in Kurmanji, etc.—;
- absence of a definiteness mark in Kurmanji implies that the machine translation system would have to add the English article;
- cases in Kurmanji might simplify the task of identifying the sentence constituents by the machine translation system when compared to Sorani.

Kurdish is a strongly inflected language. This can cause data sparseness problems if the MT system treats the words as atomic units. It is desirable that that the different grammatical suffixes are represented as independent tokens to allow the system to generalize better from the training data. Moreover, the absence of specific news-related bilingual corpora may be an obstacle to good results in a media monitoring task. The lack of a single standard in the case of both languages may result in non-homogeneous corpora.

2.10 Macedonian (mk, mkd, mac)

2.10.1 Factsheet

Macedonian is the official language of North Macedonia and also a minority language in parts of Albania, Romania, and Serbia, and it is spoken as a first language by about 2,000,000 people.

Macedonian's *endonymic glossonym*, transliterated, is *makedonski jazik*. Macedonian's closest relatives are Bulgarian (Macedonian dialects are part of a continuum with Bulgarian dialects) and, farther on, Serbo-Croatian (see sections 2.4 and 2.12). The Internet country code top-level domain for North Macedonia is `.mk`.

Macedonian uses the Cyrillic alphabet, but differs in some letters from other languages with the same script.

2.10.2 Contrasts with English

Macedonian and Bulgarian are so closely related that the contrasts described for Bulgarian in section 2.3.2 are basically valid. Minor formal differences with Bulgarian exist such as the existence in Macedonian of three forms (proximal, medial, and distal) of the definite article instead of just one (the same characteristic is only dialectal in Bulgarian). The most common form of the definite article in Macedonian (medial) is equivalent to the one used in standard Bulgarian.

2.10.3 Corpora

Monolingual corpora: Macedonian Wikipedia is available on https://mk.wikipedia.org/. As of April 2019 it contains 98,188 entries (see stats on https://stats.wikimedia.org/EN/SummaryMK.htm). A monolingual dump is periodically made available at https://dumps.wikimedia.org/mkwiki/.

This is a list of Macedonian newspapers and magazines: *Nova Makedonija* (https://www.novamakedonija.com.mk/), *Vecher* (https://vecer.mk/), *Sloboden Pechat* (https://www.slobodenpecat.mk/), *Nezavisen Vesnik* (https://nezavisen.mk/), *Kapital* (https://kapital.mk/), *Zenit* (http://zenitprilep.com.mk/).

In addition to DW (https://www.dw.com/mk/), the following international media outlets publish content in Macedonian: The Voice of America (https://mk.voanews.com/) and Radio Free Europe (https://www.slobodnaevropa.mk).

Bilingual corpora: The Southeast European Times (SETimes) is a central source of news and information about Southeastern Europe in ten languages: Albanian, Bosnian, Bulgarian, Croatian, English, Greek, Macedonian, Romanian, Serbian and Turkish. The SETimes corpus (http://www.statmt.org/setimes/) was compiled and put in the public domain by Tyers and Serdar Alperen (2010). The Macedonian–English corpus contains approximatedly 153,000 sentences.[16]

ParaSol (http://parasolcorpus.org/) is a parallel aligned corpus of translated and original belletristic texts in Slavic and some other languages. The amount of parallel corpora depends on the particular language pair. Languages include Bulgarian, Belarusian, Czech, Croatian, Macedonian, Polish, Russian, Slovak, Slovene, Serbian, Ukrainian, Upper Sorbian, German, English, Dutch,

[16]There is a cleaner version available at http://nlp.ffzg.hr/resources/corpora/setimes/.

corpus	doc's	sent's	en tokens	mk tokens
OpenSubtitles v2018	3948	3.7M	28.9M	23.7M
OpenSubtitles v2016	2966	2.9M	22.4M	18.1M
OpenSubtitles v2012	918	1.0M	8.2M	6.5M
OpenSubtitles v2013	922	1.0M	8.0M	6.4M
SETIMES v2	1	0.2M	5.1M	5.2M
SETIMES v1	1	0.2M	4.3M	4.5M
GNOME v1	1030	0.5M	2.3M	2.5M
OpenSubtitles v2011	248	0.3M	2.0M	1.7M
GlobalVoices v2017q3	2292	50.0k	1.2M	1.1M
GlobalVoices v2015	2197	46.9k	1.1M	1.0M
KDE4 v2	450	85.0k	0.5M	0.4M
Ubuntu v14.10	171	42.0k	0.3M	0.2M
EUbookshop v2	13	2.6k	0.1M	96.3k
total	15157	10.0M	84.5M	71.4M

Table 9: Distribution of the sentences in the Opus Macedonian–English corpus.

Spanish, French, Italian, and a few others. Macedonian texts are tagged and lemmatized. The Macedonian part has 1,193,788 tokens and 49,678 lemmas, whereas the English part has 814,289 tokens and 19,886 lemmas. Access to ParaSol and downloads are provided by a web interface which requires authentication.

The novel "1984" by George Orwell tagged with lemma and PoS in Macedonian and English can be downloaded from https://www.clarin.si/repository/xmlui/handle/11356/1043. English original has 79,718 sentences and 106,4424 words. The corpus is licensed under a CC BY-NC-SA 4.0 license.

Opus (http://opus.nlpl.eu) has a Macedonian–English corpus of approximately 10 million sentences distributed as shown in table 9.

Google Translate supports Macedonian–English translation. Apertium (see below) also has Macedonian–English translation as well as Macedonian–Serbo-Croatian, and to a lesser extent Macedonian–Albanian and Macedonian–Bulgarian.

2.10.4 Resources

Monolingual resources: An Crúdabán (http://crubadan.org/writingsystems) has lists of words and bigrams for different variants and scripts of Kurdish; files containing a small number of URLs that were used to compile the words are also included.

Apertium includes stable linguistic data for Macedonian–English and Macedonian–Serbo-Croatian in the form of morphological analysers, bilingual dictionaries and rule-based machine translation. The language pairs Macedonian–Albanian and Macedonian–Bulgarian seem to be in an earlier stage of development. The annoucement in 2010 of the first release of the Macedonian–English system (see https://www.mail-archive.com/apertium-stuff@lists.sourceforge.net/msg00271.html) stated that it contained around 8,000 dictionary entries and 66 transfer rules. The data can be downloaded from https://github.com/apertium?q=mk and is published under a GPL2 or GPL3 license

depending on the language pair.

The MULTEXT-East non-commercial Macedonian lexicon (https://www.clarin.si/repository/xmlui/handle/11356/1042) contains 1,323,572 entries with surface form, lemma and morphosyntactic tags.

Bilingual resources: A crowd-sourced Macedonian–English dictionary (http://www.seas.upenn.edu/~nlp/resources/TACL-data-release/dictionaries.tar.gz) has been made available by Pavlick et al. (2014).

The so-called machine readable English–Macedonian dictionary (https://time.mk/trajkovski/tools/dict/) contains 23,296 translation pairs. The dictionary is provided under the Creative Commons Attribution-NonCommercial 3.0 Unported License and its development has been described by Saveski and Trajkovski (2010).

See the description in the previous section about the resources provided by the Apertium project.

2.10.5 Challenges for corpus-based MT from English

As already stated, the linguistic contrasts are basically the same as those between Bulgarian and English (see section 2.3.2) and so are the challenges for corpus-based MT from English (see section 2.3.5). The fact that, unlike Bulgarian, Macedonian is not an official language of the European Union results in a smaller availability of Macedonian-English parallel corpora, but the high degree of similarity between Macedonian and Bulgarian becomes an opportunity as transfer learning from an NMT system trained from English to Bulgarian may work with Macedonian.

Note that language detection systems are unlikely to misclassify Macedonian as Bulgarian (or vice versa), as there are a few different letters in the Cyrillic alphabets of both languages: for example, the sound /dz/ is represented by the letter Dze (which may look identical to the Latin letter S) in Macedonian and a digraph in Bulgarian.

2.11 Punjabi (pa, pan)

2.11.1 Factsheet

According to Wikipedia, Panjabi or Punjabi has around 120 million speakers. It is spoken (and official) in the Pakistani province of Punjab and in India, in the states of Punjab, Haryana, the Union territory of Chandigarh and the capital city, Delhi. It belongs to a very large family, the Indo-European family, as do many other languages in this project, namely Bosnian (§ 2.2), Croatian (§ 2.4), Serbian (§ 2.12), Bulgarian (§ 2.3), and Macedonian (§ 2.3); however, these five languages are Slavic, while Gujarati (§ 2.5) belongs to the Indo-Arian group.

Punjabi is written both in the Shahmukhi alphabet (based on the Perso-Arabic script, mainly in Pakistan) and in the Gurmukhi *alphasyllabary* in which consonant–vowel groups are written as a single character, with vowels being secondary to consonants.

2.11.2 Contrasts with English

What follows are tables of contrasts between English and Punjabi. They are not meant to be exhaustive (many other contrasts are described in WALS[17]). Examples (where available) are given in Google-style transliteration.

Function words			
Feature	**Value in English**	**Value in Punjabi**	**Examples**
Indefinite article	Different from 'one'	No indefinite article	*oha kute nala a'i'a* 'They came with a small dog', lit. 'They small dog with came'
Definite article	Definite word distinct from demonstrative	No definite article	*adamı* 'The man' or 'A man'.

[17] https://wals.info/

Morphology			
Feature	**Value in English**	**Value in Punjabi**	**Examples**
Politeness distinction in pronouns	No distinction	Binary distinction	
Position of case affixes	No case affixes, no pre- or post-positional clitics	Case suffixes	*Ghara* '[The] house'. *Ghara vica* 'In [the] house'. *Ghara toṁ* 'From the house'.
Number of genders	Three (sex-based, only in singular pronouns and possessives)	Two (formally and semantically assigned)	*caga ghara* 'good house' (masc.) but *cagı premıka* 'good girlfriend' (fem.)

Syntax			
Feature	**Value in English**	**Value in Punjabi**	**Examples**
Order of Subject, Object and Verb:	Subject–Verb–Object	Subject–Object–Verb	*Adamı ne ghara kharídi'á* 'The man bought the house', lit. 'Man (particle) house bought'
Postpositions or prepositions?	Prepositions	Postpositions	*Ghara de piche* 'behind the house', lit. 'House-of back'; *Ghara de sahamaṇe* 'In front of the house', lit. 'House-of front'
Order of genitive and noun	No dominant order	Genitive–noun	*Ghara dı chata* 'The roof of the house', lit. 'House of roof'.
Polar questions	No question particle, reordering and optional use of auxiliaries	Initial question particle, no reordering	*Bilı kala hai* 'The cat is black' *kı bilı kala hai?* 'Is the cat black?'
Position of interrogative phrase in content questions	Initial	Non-initial	*Tusıṁ* ***ghara*** *vekhade ho* 'You see **houses**'; *Tusıṁ* ***kı*** *vekhade ho?* '**What** do you see?'
The transitive verb agrees with...	...the A argument ('agent'), if it does	...the A argument in imperfective verb forms, and the P argument ('patient') in perfective verb forms (split ergativity) [18]	*Adamı ne kara kharıda la'i* 'The man bought the car', *Adamı ne kara kharıda la'e* 'The man bought the cars' (note that the verb agrees with the P argument in the perfect).

2.11.3 Corpora

Monolingual corpora: There are a number of monolingual corpora available as part of wikimedia dumps: the Punjabi Wikipedia contains 31,145 articles (https://dumps.wikimedia.org/pawiki/; Punjabi Wiktionary contains 13,795 entries (https://dumps.wikimedia.org/pawiktionary/); Punjabi Wikibooks cotains 64 books (https://dumps.wikimedia.org/pawikibooks/); Punjabi Wikisource consists of 200 texts (https://dumps.wikimedia.org/pawikisource/).

There are media outlets, in addition to BBC, containing Punjabi content which could be crawled to obtain monolingual corpora: TWR360 (https://www.twr360.org/), although most of it is multimedia content; Punjab Infoline News Network (http://punjabi.punjabinfoline.com/), and Punjabi

[18] Also shown by Gujarati, see section 2.5 for more details.

corpus	doc's	sent's	en tokens	pa tokens
GNOME v1	1,428	500k	3.0M	3.5M
KDE v4	999	99k	0.8M	0.4M
Ubuntu 14.10	221	63.9k	0.3M	0.3M
total	2,648	662.9k	4.1M	4.2M

Table 10: OPUS resources for Punjabi–English.

Tribune (https://www.punjabitribuneonline.com/).

Other sites from which monolingual corpora could be crawled: the Universal Declaration of Human Rights (http://unicode.org/udhr/d/udhr_pan.html) and the Bible (https://live.bible.is/bible/PANWTC/MAT/1)

Bilingual corpora: The amount of parallel corpora freely available is very scarce. Table 10 reports the amount of parallel sentence and words in each language for the corpora available at OPUS (http://opus.nlpl.eu/).

The EMILLE/CIIL Corpus (http://catalog.elra.info/en-us/repository/browse/ELRA-W0037/) contains bilingual corpora for Punjabi as well as for other language of interest to the GoURMET (Gujarati).

The website Jehovah's Witnesses (https://www.jw.org), could be crawled to obtained Punjabi–English parallel corpora in the religious domain.

Jindal et al. (2017) describes an English–Punjabi parallel corpus and cites several sources used to build the corpus, but the corpus does not seem to be available.

2.11.4 Resources

Monolingual resources: There is a open-source Punjabi grammar developed in Grammatical Framework (Virk et al. (2011); http://www.grammaticalframework.org/lib/src/punjabi/).

Bilingual resources: A crowd-sourced Punjabi–English dictionary (http://www.seas.upenn.edu/~nlp/resources/TACL-data-release/dictionaries.tar.gz) has been made available by Pavlick et al. (2014). It contains 98,027 bilingual entries.

There are also three online bilingual dictionaries: Glosbe (https://glosbe.com/pa/en/), Shabdkosh (https://www.shabdkosh.com/dictionary/english-punjabi/) and The Panjabi Dictionary (http://dsal.uchicago.edu/dictionaries/singh/).

Finally, Google Translate offers MT between Punjabi and English.

2.11.5 Challenges for corpus-based MT to English

Here are the main challenges when generating English from Punjabi:

- Scarcity of bilingual corpora.
- The order of genitive constructions maps differently.

- The absence of definite and indefinite articles in Punjabi may make the generation of grammatical English difficult.
- Extensive reordering (bringing the verb forward, or generating prepositions) may be challenging for complex sentence structures.

2.12 Serbian (sr, srp)

2.12.1 Factsheet

According to Wikipedia, Serbian is the standardized variety of the Serbo-Croatian language mainly used by 9–10 million Serbs. "It is the official language of Serbia, co-official in [...] Kosovo, and one of the three official languages of Bosnia and Herzegovina. In addition, it is a recognized minority language in Montenegro where it is spoken by the relative majority of the population as well as in Croatia, North Macedonia, Romania, Hungary, Slovakia, and the Czech Republic."

Serbian uses both the same Latin alphabet as Croatian (§ 2.4) and a Cyrillic alphabet similar to that of Russian but containing some special letters corresponding to Croatian *ć đ*, *dž*, *j*, *lj* and *nj*. Most Serbian speakers are used to both alphabets, and transliteration is basically one-to-one either way.

2.12.2 Contrasts with English

Serbian has essentially the same contrasts with English as Croatian, see section 2.4.2; there are, however, some differences:

- differences in vocabulary (for instance, Serbian *hleb* instead of Croatian *kruh* for 'bread' or Serbian *januar*, *februar*, ... instead of Croatian *siječanj*, *veljača*, ...);
- small differences in grammar (for instance Serbian *Želim da znam* 'I want to know', lit. 'I want that I know' instead of Croatian *Želim znati* 'I want to know', also lit.;
- there are a few spelling differences between Serbian and Croatian: for instance, Serbian often has *e* (*mleko* 'milk', *razumela je* 'She understood') where Croatian has *ije* (*mlijeko*) or *je* (*razumjela je*).

2.12.3 Corpora

Bilingual corpora: The Southeast European Times (SETimes) is a central source of news and information about Southeastern Europe in ten languages: Albanian, Bosnian, Bulgarian, Croatian, English, Greek, Macedonian, Romanian, Serbian and Turkish. The SETimes corpus (http://nlp.ffzg.hr/resources/corpora/setimes/) was compiled and put in the public domain by Tyers and Serdar Alperen (2010) and refined by the Natural Language Processing group at the University of Zagreb. The Serbian–English corpus contains approximately 225,000 sentences.

ParaSol (http://parasolcorpus.org/) is a parallel aligned corpus of translated and original belletristic texts in Slavic and some other languages. The amount of parallel corpora depends on the particular language pair. Languages include Bulgarian, Belarusian, Czech, Croatian, Macedonian, Polish, Russian, Slovak, Slovene, Serbian, Ukrainian, Upper Sorbian, German, English, Dutch, Spanish, French, Italian, and a few others. Croatian texts are tagged and lemmatized. The Serbian part has 1,324,929 tokens and 42,602 lemmas, whereas the English part has 814,289 tokens and 19,886 lemmas. Access to ParaSol and downloads are provided by a web interface which requires authentication.

The novel "1984" by George Orwell tagged with lemma and part-of-speech in Serbian and English can be downloaded from https://www.clarin.si/repository/xmlui/handle/11356/1043. English

corpus	doc's	sent's	en tokens	sr tokens
OpenSubtitles v2018	55,422	45.9M	371.3M	301.0M
OpenSubtitles v2016	28,398	24.2M	194.5M	157.2M
OpenSubtitles v2012	16,325	15.5M	124.9M	100.3M
OpenSubtitles v2013	15,228	14.5M	117.6M	94.2M
GNOME v1	1,547	0.6M	3.2M	3.5M
GlobalVoices v2017q3	1,023	20.1k	0.7M	0.4M
Ubuntu v14.10	421	97.0k	0.7M	0.3M
KDE4 v2	764	64.5k	0.5M	0.5M
GlobalVoices v2015	932	17.1k	0.6M	0.4M
total	120060	100.3M	817M	657M

Table 11: Distribution of the sentences in the Opus Serbian–English corpus.

original has 79,718 sentences and 106,4424 words. The corpus is licensed under a CC BY-NC-SA 4.0 license.

The srenWaC — Serbian-English Parallel Web Corpus (https://www.clarin.si/repository/xmlui/handle/11356/1059/) consists on 534,682 Serbian–English sentence pairs. It is published under the CC-BY-SA license.

The ParCoLab French-Serbian-English corpus (http://parcolab.univ-tlse2.fr/en/about/resources/) contains two Serbian–English parallel subcorpora: the first is extracted from the Web magazine Pescanik (it contains 31,151 Serbian tokens and 34,275 English tokens) and the second one from TED talks (it contains 18,933 Serbian tokens and 21,410 English tokens).

Opus (http://opus.nlpl.eu) has a Serbian–English corpus of approximately 110 million sentences distributed as shown in table 11.

Monolingual corpora: srWaC is a web corpus collected from the .sr top-level domain by Ljubešić and Klubička (2014). It contains 894 million tokens and is annotated with the lemma, morphosyntax and dependency syntax layers.

The Twitter corpus of BCMS (Ljubešić et al., 2014) contains 379,255,987 words in Bosnian/Croatian/Montenegrin/Serbian. It is distributed as a set of tweet ids that should be used to rebuild the corpora via the Twitter API.

The Serbian Wikipedia is available at https://sr.wikipedia.org/wiki/. As of April 2019 it contains 614,211 entries (see stats on https://stats.wikimedia.org/EN/SummarySR.htm). A monolingual dump is periodically made available at https://dumps.wikimedia.org/srwiki/. There is a Wikipedia dump already preprocessed that is available in plain text format (http://hdl.handle.net/11234/1-2735). It contains around 25 million sentences and 66 million words.

The W2C (Web to Corpus) corpora is a set of corpora (https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0022-6133-9) for 120 languages automatically collected from Wikipedia and the web. The Serbian corpus contains around 250,000 sentences and 18 million words.

In addition to BBC, both in Latin alphabet (https://www.bbc.com/serbian/lat) and in Cyrillic (https://www.bbc.com/serbian/cyr), and DW (https://www.dw.com/sr), the following international media outlets produce content in Serbian: Global Voices (https://sr.globalvoices.org/, Latin), The

Voice of America (https://www.glasamerike.net/, Latin) , China Plus (formerly China Radio International, http://serbian.cri.cn , Latin), and TWR360 (https://www.twr360.org/, Cyrillic), although mostly multimedia content). These outlets may be interesting sources from which to obtain monolingual corpora.

2.12.4 Resources

Bilingual resources: Apertium provides rule-based machine translation between Bosnian/Croatian/Serbian and English. The rule-based system contains bilingual dictionaries and transfer rules released under free licenses (https://github.com/apertium/apertium-hbs-eng).

PanLex contains a bilingual Serbian–English dictionary that can be queried online (https://translate.panlex.org/?langDe=eng-000&langAl=srp-000).

A crowd-sourced Serbian–English dictionary (http://www.seas.upenn.edu/~nlp/resources/TACL-data-release/dictionaries.tar.gz) has been made available by Pavlick et al. (2014).

The set of open-source bilingual dictionaries FreeDict (https://github.com/freedict/fd-dictionaries) contains an Serbian–English bilingual dictionary.

The Glosbe bilingual concordancer can be used online at https://glosbe.com/en/sr.

The following online machine translation systems support Serbian–English translation:

- Google Translate (https://translate.google.com)
- Yandex Translate (https://translate.yandex.com)
- Bing Translator (https://www.bing.com/translator)

Monolingual resources Apertium contains a morphological analyser/part-pf-speech tagger/-morphological generator for Bosnian, Croatian and Serbian (https://github.com/apertium/apertium-hbs) with 58,004 stems. The srLex morphological lexicon (Ljubešić et al., 2016), released under a free licence, can be used together with a conditional random fields tagger (Ljubešić and Erjavec, 2016) to morphologically analyze Serbian text.

Wikimorph-sr (http://redac.univ-tlse2.fr/lexiques/wikimorph-sr_en.html) is a morphosyntactic lexicon for Serbian that can be used for part-of-speech tagging, parsing and lemmatisation. It was mainly extracted from the Serbo-Croatian edition of the Wiktionary (sh.wiktionary.org).

2.12.5 Challenges for corpus-based MT from English

The linguistic contrasts between English and Serbian are basically the same as those between English and Croatian (§ 2.4.2), and so are the challenges for corpus-based MT from English (§ 2.4.5). The fact that Serbian can be written in the Cyrillic alphabet is not a relevant issue, since it can be transliterated from the Latin alphabet with a one-to-one mapping.

In addition, the main problem lies in harvesting new corpora for Serbian, as language identification is very likely to (expectedly) classify Croatian or Bosnian text as Serbian written with the Latin alphabet or vice-versa.

2.13 Swahili (sw, swa)

2.13.1 Factsheet

According to Wikipedia, Swahili (also called *Kiswahili*) has estimates of between 2 and 15 million first-language speakers and about 90 million second-language speakers. It is spoken in Tanzania*, Democratic Republic of the Congo*, Kenya, Somalia (on the Bajuni islands and Barawa), Mozambique (mostly Mwani), Burundi, Rwanda*, Uganda*, Comoros, Mayotte, Zambia, Malawi, and Madagascar (official where marked with *). It belongs to a very large family, the Niger–Congo family, as Yoruba (§ 2.16) and Igbo (§ 2.7), and more specifically to the Bantu group. Swahili is currently written in the Latin script, with no diacritics; the apostrophe is used in the seldom-occurring combination *ng'* which represents the sound of *ng* in *singer* (not *finger*), and can occur at the beginning of a word (*ng'ombe*, 'cow')

2.13.2 Contrasts with English

The following tables summarize the main contrasts between Swahili and English. Some examples are from Perrott (1965).

Nouns			
Feature	**Value in English**	**Value in Swahili**	**Examples**
Coding of plurality in nouns	Plural suffix	Plural prefix	***ki**chwa* ('head'), ***vi**chwa* ('heads'); ***ji**cho* ('eye'), ***ma**cho* ('eyes')

Verb			
Feature	**Value in English**	**Value in Swahili**	**Examples**
Number of categories encoded in a single-word verb	Few (number, person, tense)	Many ("STROVE", that is, number and person of **s**ubject, **t**ense, aspect and mood, optional **r**elatives, number and person of **o**bject, **v**erb root, and optional **e**xtensions)	*nimekinunua kitabu* 'I have bought the book', where: *ni* 'I', subject; *me*, present perfect; *ki*, 'it', object; *nunua*, 'buy', verb root.

Function words			
Feature	**Value in English**	**Value in Swahili**	**Examples**
Definite articles	Definite word distinct from demonstrative	Demonstrative (seldom) used as definite article	*kitabu* ('book', 'the book', 'a book').
Noun Phrase Conjunction	*And* different from *with*	*And* identical to *with*	*Lete chai na maziwa* ('Bring tea and milk'); *Yesu alikuja na Baba yake* ('Jesus came with his Father').

Morphology			
Feature	**Value in English**	**Value in Swahili**	**Examples**
Inflectional morphology	Suffixing	Mainly prefixing	*kitabu* ('book'), *vitabu* ('books'); *nilinunua* ('I bought'), *ulinunua* ('You bought'); but *jenga* ('build'), *jengwa* ('be built')

Syntax			
Feature	**Value in English**	**Value in Swahili**	**Examples**
Reduplication	No productive reduplication	Productive full and partial reduplication	*nikaenda* 'I went'; *nikaenda nikaenda* 'I went on (and on)'
Number of genders	Three, sex-based, only in 3rd person singular pronouns and possessives	Many, not based on sex (called *classes*)	*kitabu* 'book' (*ki-vi*-class): plural *vitabu* 'books' ; *mtoto* 'child' (*m-wa*-class): plural *watoto* 'children' ; etc. Note that adjectives and verbs have to agree: *kitabu kidogo* 'small book', *vitabu vidogo* 'small books'; *mtoto mdogo* 'small child', etc.
Order of genitive and noun	No dominant order	Noun–genitive	*gari la mama* 'Mom's (*mama*) car (*gari*)'; *paa la nyumba* 'The roof (*paa*) of the house (*nyumba*)'.
Order of adjetive and noun	adjective–noun	noun–adjective	*mtoto mdogo* 'small child', lit. 'child small'
Order of demonstrative and noun	demonstrative–noun	noun–demonstrative	*gari hili* 'this car', lit. 'car this'
Order of numeral and noun	numeral–noun	noun–numeral	*vitabu viwili* ('two books', lit. 'books two')
Expression of Pronominal Subjects	Obligatory pronouns in subject position	Subject affixes on verb	*Nilinunua* ('I bought'), *ulinunua* ('You bought')
Negation	Particle or construction	Negative form of verb	*Ninasoma* ('I read'), *Sisomi* ('I do not read'); *Unasoma* ('You read'), *husomi* ('You do not read');
Position of Interrogative Phrases in Content Questions	Initial interrogative phrase	Not initial interrogative phrase	*Unasoma vitabu* ('You read books'); *Unasoma nini?* ('What do you read', lit. 'you read what?')
Polar questions	Change in word order, use of auxiliaries	No change in word order	*Amesoma* ('He has read'); *Amesoma?* ('Has he read?')
Comparative	Comparative form of adjective ('-er') or 'more'	Absolute form of adjective	*Virusi ni ndogo* ('A virus is small') *Virusi ni ndogo kuliko bakteria* ('A virus is smaller than a bacterium', lit. 'A virus is small where there is a bacterium')
Predicative Possession	'have'	conjunctional ('to be with')	*Nina swali* ('I have a question', lit. 'I-am-with question')

corpus	doc's	sent's	en tokens	sw tokens
Tanzil (Quran)	15	0.1M	2.8M	2.1M
GlobalVoices (v2015)	1,349	26.1k	0.6M	0.6M
GlobalVoices (v2017q3)	1,474	29.7k	0.7M	0M
Ubuntu	45	1.0k	59.6k	3.6k
EUBookshop v2	3	17	0.3k	0.3k
GNOME v1	3	40	0.3k	0.2k
total	2,889	0.2M	4.2M	2.7M

Table 12: OPUS resources for Swahili–English.

2.13.3 Corpora

Monolingual corpora: Crawls of Swahili news text are available at http://data.statmt.org/news-crawl/sw/.

The Helsinki Corpus of Swahili v.2 (https://www.kielipankki.fi/news/hcs-a-v2-in-korp/) is available in unannotated form at https://korp.csc.fi/download/HCS/na-v2/hcs-na-v2.zip, and in annotated form at https://korp.csc.fi/download/HCS/a-v2/hcs-a-v2-dl. The annotations are in a format called VRT and contain: form, lemma, part of speech, morphological indicators, *English translation*, dependency relations, etc. The licence for the unannotated corpus is very permissive;[19] the licence for the annotated form restricts it to academic and research use.[20]

The monolingual dump of the Swahili wikipedia is periodically made available at https://dumps.wikimedia.org/swwiki/. As of April 29th it contains around 49,592 articles.

In addition to BBC and DW, the following international media outlets produce content in Swahili: Radio France Internationale (http://sw.rfi.fr/), Global Voices (https://globalvoices.org/), The Voice of America (https://www.voaswahili.com/) , China Plus (formerly China Radio International, http://swahili.cri.cn/), Vatican Radio (https://www.vaticannews.va/sw.html), and TWR360 (https://www.twr360.org/, although mostly multimedia content). These outlets may be interesting sources from which to obtain monolingual corpora.

Bilingual corpora: OPUS[21] contains ready-made corpora for Swahili (see table 12) and reports a total of 0.2 million segments.

De Pauw et al. (2011) report a Swahili–English parallel corpus (SAWA) containing 73,700k. Alacant has contacted the authors and they have made it available to GoURMET under a research-only license. The corpus provided contains about 200,000 sentence pairs, which may contain segments which are already in OPUS.

2.13.4 Resources

Monolingual resources The Unimorph project (https://github.com/unimorph) contains a set of morphologically-annotated Swahili forms (https://github.com/unimorph/swc); the coverage of

[19] http://creativecommons.org/licenses/by/4.0/
[20] https://www.kielipankki.fi/lic/hcs-a-v2-dl/?lang=en
[21] http://opus.nlpl.edu

the 2018 version of the Swahili news crawls described above is low (15%) with 10200 forms.

After extracting the unique forms (all of which are analysed morphologically) in the Helsinki Corpus of Swahili, they give a better coverage of the Swahili news crawls of about 88%.

There is a free/open-source Swahili verb segmenter in PHP: https://github.com/donnekgit/segmenter.

Lipps (2011) describes a morphological analyser for Swahili. The paper contains listings, but they are clearly marked as research-only after requesting for permission.

Bilingual resources: A crowd-sourced Swahili–English dictionary (http://www.seas.upenn.edu/~nlp/resources/TACL-data-release/dictionaries.tar.gz) has been made available by Pavlick et al. (2014).

There is a commercial bilingual Swahili–English dictionary that may be installed in Windows (https://africanlanguages.com/swahili/)

Both Google Translate and Yandex have Swahili translators.

2.13.5 Challenges for corpus-based MT to English

Swahili is morphologically and syntactically quite different from English, in spite of the fact that both are subject–verb–object languages. Swahili verb morphology is rich and agglutinative, and a large number of morphologically-marked nominal genders participate in nominal and verbal agreement.

In summary:

- Parallel corpora are rather scarce, and it remains to be seen how much more bilingual text may be crawled using tools such as Bitextor.[22] Approaches based on monolingual corpora (Artetxe et al., 2018) may be worth trying as monolingual corpora beyond a few million words may be easily retrievable. Another approach worth trying would be synthetic bilingual corpora (perhaps back-translated using commercial systems such as Google translate and Yandex).
- A correct segmentation of verbs, which are very rich and complex, is not too complicated, but it may be needed as a preprocessing step.
- Word-order differences seem to occur locally (basically inside the noun phrase). This may only be a problem for longer noun phrases.
- The absence of definite and indefinite articles in Swahili may make the generation of grammatical English tricky.
- Genders in Swahili do not mark sex; generating the correct English 3rd-person pronouns and possessives may be challenging.
- Swahili interrogatives have to be reordered when translating to English.

[22] https://github.com/bitextor/bitextor/

2.14 Tigrinya (`ti`, `tir`)

2.14.1 Factsheet

According to Wikipedia, Tigrinya has around 7 million speakers. It is spoken (and official) in Eritrea, and a recognized minority language in Ethiopia (Tigray state). It is a Semitic language.

Tigrinya is written in the Ge'ez script, an *abugida* (syllabary) which is also used for its larger neighbour Amharic (spoken in Ethiopia).

2.14.2 Contrasts with English

Examples are not provided for all contrasts described in the World Atlas of Linguistic Studies, and the tables may still miss important contrasts. Where provided, they are given in transliteration, and the main source is Wikipedia.

Syntax			
Feature	**Value in English**	**Value in Tigrinya**	**Examples**
Order of Subject, Object and Verb	Subject–Object–Verb	Object–Verb–Subject	
Adpositions: prepositions or postpositions?	Prepositions	Both	
Polar questions	Interrogative word order	Same order as affirmative	
Order of genitive and noun	No dominant order	Noun–genitive	
Negative Morphemes	Negative particle	Negative affix	*yəsəbbäruläy* ('they are broken for me'), *ayyəsəbbäruläyə* ('they are not broken for me')
Passive constructions	Present	Absent	

Syntax			
Feature	**Value in English**	**Value in Tigrinya**	**Examples**
Prefixing vs. Suffixing in Inflectional Morphology	Strongly suffixing	Equal prefixing and suffixing	
Possessives	Separate words	Affixes	*gäza* 'house', *gäza-y* 'my house', *gäza-'a* 'her house'; also for pronouns after prepositions *bəza'ba* 'about', *bəza'ba-y* 'about me', *bəa'bə-'a* 'about her'
Noun inflection	Mainly suffix (plural)	Semitic template and suffix	Template: *färäs* 'horse', *'afras* 'horses'. Suffix: *'arat* 'bed', *'aratat* 'beds'.
Verbal Person Marking	Only the A argument (agent)	A (agent) or P (patient) argument	*rə'yä-yya* ('I-saw-her', both the subject and object are marked as part of the verb)
The Morphological Imperative	No specific second-person imperatives	Second singular and second plural	
Gender distinction in pronouns	Only 3rd singular	2nd and 3rd, singular and plural	

2.14.3 Corpora

Monolingual corpora: There are corpora available as part of wikimedia dumps: Tigrinya Wikipedia contains 169 articles (https://dumps.wikimedia.org/tiwiki/ and Tigrinya Wiktionary contains 115 entries (https://dumps.wikimedia.org/tiwiktionary/).

The Nagaoka Tigrinya Corpus (Tedla et al., 2016) (https://eng.jnlp.org/yemane/ntigcorpus) consists of news articles and contains 72,080 tokens annotated with part-of-speech. The corpus is encoded in TEI.

The Tigrinya Web Corpus (tiWaC; http://hdl.handle.net/11234/1-2592) is available for research after login with an account from an European university. Information about the web pages from which the texts were crawled is provided at https://habit-project.eu/wiki/TigrinyaCorpus. The corpus contains part-of-speech annotations based on Universal dependencies.[23]

In addition to BBC, the following media outlets produce content in Tigrinya: the Voice of America (tigrinya.voanews.com), Vatican Radio (https://www.vaticannews.va/ti.html) and Asmarino News (http://www.asmarino.com/tig). Eritrea Haddas publishs PDFs of weekly newspapers published by the Eritrean Ministry of Information (http://www.shabait.com/eritrea-haddas).

Other sources from which monolingual corpora could be downloaded include the Universal Declaration of Human Rights (http://unicode.org/udhr/d/udhr_tir.html) and the URLs included in the Crúbadán repository for Tigrinya (http://crubadan.org/languages/ti).

[23] https://www.sketchengine.eu/universal-pos-tags/

Bilingual corpora: OPUS[24] contains very little bilingual material for Tigrinya: it reports 1,400 segment pairs of Ubuntu translations with 42,300 words in English and 2,000 word in Tigrinya. The Tigrinya side of the corpus is plenty of English text.

What follows is a list of websites from which parallel corpora could be harvested, all in the religious domain: Jehovah's Witnesses (https://www.jw.org), Ethiopicbible (https://www.ethiopicbible.com), Ebible (http://ebible.org) and Ge'ez experience (https://www.geezexperience.com).

2.14.4 Resources

Monolingual resources: HornMorpho (https://github.com/adamsamson/HornMorpho2.5) is a morphological analyser and generator for Tigrinya, as well as for Amharic and Afaan Oromoo (§ 2.1).

Bilingual resources: The GeezLab Tigrinya BiLingual Lexicon (https://github.com/fgaim/Tigrinya-BiLexicon) contains 78642 bilingual entries statistically obtained from a parallel corpus that the authors do not specify.

There are three online bilingual dictionaries —Tigrinya dictionary (https://www.geezexperience.com/), Memhr.org Dictionary (urlhttp://www.memhr.org/dic/) and Glosbe (https://glosbe.com/ti/en)—. and a phrasebook available for learners of Tigrinya as a foreign language (http://www.goethe-verlag.com/book2/EN/ENTI/ENTI002.HTM).

Finally, to our knowledge, no online commercial MT systems offers Tigrinya.

2.14.5 Challenges for corpus-based MT from English

Here are the main challenges when generating Tigrinya from English:

- Scarcity of bilingual corpora.
- Radically different sentence structure. In particular, interrogatives have to be reordered when translating from English.
- The absence of passive voice may make the translation of the English passive voice difficult.
- The absence of gender distinction on second person pronouns in English may make difficult to translate to the appropriate second-person pronoun in Tigrinya.

[24] http://opus.nlpl.edu

2.15 Turkish (tr, tur)

2.15.1 Factsheet

According to Wikipedia, Turkish has around 75 million first-language (L_1) speakers and 85 million who speak it either as L_1 or L_2. It is spoken (and official) in Turkey, Northern Cyprus, and Cyprus and it is a recognized minority language in Bosnia and Herzegovina, Greece, Iraq, Kosovo, Macedonia and Romania. It is the largest language in the Turkic language family, followed by Azeri, Uzbek, Kazakh and Uygur.

Turkish is written in a modified Latin script that contains some letters with diacritics: *ç*,*Ç*, *ğ*, *Ğ*, *ı*, *İ*, *ö*, *Ö*, *ş*, *Ş*. The letters *â*/*Â*, *î*/*Î* and *û*/*Û* are also used occasionally to mark palatalization of the following consonant or lengthening of the vowel. The script was introduced and adopted by the Turkish republic in the late 1920's and is widely and consistently used.

Unlike English, Turkish is an agglutinative language (it creates long words made up of many morphemes) with extensive vowel harmony and some consonant harmony (vowels and consonants to either side of a morpheme boundary have to belong to the same family) Each morpheme usually distinctly encodes a single category, an exception being the fusion of person and number in pronouns, possessives and verbs.

2.15.2 Contrasts with English

Many examples taken from https://en.wikipedia.org/wiki/Turkish_grammar.

Verbs			
Feature	**Value in English**	**Value in Turkish**	**Examples**
Number of categories encoded in a single verb form	A few (person, number, tense)	Many (also potentiality, negation, voice [passive/active], evidentiality ['it seems'], etc.)	*gelmişim* ('It seems that I came') vs *geldim* ('I came')
Perfective/ imperfective aspect	No grammatical marking	Grammatical marking	*yürüdüm* (≃ 'I walked [and had finished walking]') vs. *yürüyordum* (≃ 'I walked [and had not finished walking]')
Possibility (situational or epistemic)	Verbal constructions	Affixes on verbs	*Gelebirim* ('I can come'), *Gelemem* ('I cannot come')
Evidentiality	Expressed through constructions	Morphologically expressed	*O gelmiş* ('It seems that she came') vs. *O geldi* ('She came').
Negative Morphemes	Negative particle	Negative affix	*Anladım.* ('I understood'); *Anlamadım* ('I did not understand').

Morphology			
Feature	**Value in English**	**Value in Turkish**	**Examples**
How is case expressed	It is not	It is by a single, distinct morpheme	Case in nouns is expressed by a morph that comes just after the number and possessive morphs: *Bahçelerimde* ('in my gardens, lit. *Bahçe* 'garden' *ler* plural mark, *im* first person possessive mark, *de* locative 'in').
Number of morphologically-marked cases	Two (but only in pronouns, etc.)	Six (nominative, accusative, dative, locative, ablative and genitive), appearing as distinct suffixes.	• *Ev yandı* ('the house burned down', nominative, no mark); • *Ev**i** gördüm* ('I saw the house', accusative *-i*); • *Ev**e** gittim* ('I went to the house', dative *-e*); • *Ev**de** yattım* ('I slept in the house, locative *-de*); • *Ev**den** geldim* ('I came from the house, ablative *-den*); • *Ev**in** çatısı* ('The roof of the house', genitive *-in*).
Position of Pronominal Possessive Affixes	No possessive affixes	Possessive suffixes	*Ev yandı* ('[the] house burned down') vs. *Ev**im** yandı* ('**my** house burned down')

Syntax			
Feature	**Value in English**	**Value in Turkish**	**Examples**
How is possession marked?	Only on the dependent (possessor in this case).	Both at the dependent (possessor) and the head (possessed)	*başkan**ın** ev**i*** (the president's house, lit. *başkan*, 'president'; *ın*, GENITIVE; *ev*, 'house'; *i*, 3RD-PERSON-POSSESSIVE)
Reduplication	No productive reduplication	Productive full and partial reduplication	May lead to a different sense: *zaman zaman* ('occasionally', lit. 'time time')
Order of Subject, Object and Verb	Subject–Verb–Object	Subject–Object–Verb	*Peter arabayı gördü* ('Peter saw the car', lit. 'Peter the-car saw')
Order of Object and Verb	Verb–Object	Object–Verb	*Kedi **fareleri kovalamaya** geldi* ('The cat came **to chase mice**', lit. '[The] cat mice to-chase came')
Order of Object, Oblique, and Verb	Verb–Object–Oblique	Oblique–Object–Verb	*Dükkandan bir kitap aldım* ('I bought a book from the shop', lit. 'Shop-from one book bought-I')
Adpositions: Prepositions or postpositions?	Prepositions (before noun)	Postpositions (after noun)	*Dükkan* ('shop'); *Dükkandan* ('from the shop', lit. 'shop-from')
Order of Genitive and Noun	No dominant order	Genitive–Noun	*Ev**in** çatısı* ('The roof of the house', genitive *-in*).
Order of Relative Clause and Noun	Noun–Relative clause	Relative clause–Noun	*Camı kıran adamı gördüm* ('I saw the man who broke the window', lit. 'window-the broke-who man-the saw-I')
Position of Polar Question Particles	No question particle	Final	*Onu gördün* ('You saw her'); *Onu gördün mü?* ('Did you see her?')
Position of Interrogative Phrases in Content Questions	Initial	Not initial	*Onu nerede gördün?* ('Where did you see her', lit. 'Her where saw-you?')

Function words			
Feature	**Value in English**	**Value in Turkish**	**Examples**
Definite article	Yes, different from demonstrative.	No definite article	*başkan* may mean 'president' or 'the president' (note, however, that when the noun is an object, the presence of the accusative case ending may have a similar function as English 'the' *Adam kapı**yı** kapatt* ('The man closed the door', lit 'Man door-ACCUSATIVE closed').
Indefinite article	Different from 'one'	Same word as 'one'	*Küçük bir evim var* may be interpreted as 'I have one small house' and 'I have a small house'

Pronouns			
Feature	**Value in English**	**Value in Turkish**	**Examples**
Politeness distinction in pronouns	No distinction	Binary politeness distinction	*sen* is the familiar 2nd person singular pronoun; the polite form is *siz*, which is the same as the 2nd person plural.
Expression of Pronominal Subjects	Obligatory pronouns in subject position	Subject affixes on verb	*Geldim* ('I arrived'); *geldimiz* ('We arrived').

2.15.3 Corpora

Monolingual corpora: Crawls of Turkish news text are available at http://data.statmt.org/news-crawl/tr/.

The monolingual dump of the Turkish wikipedia is periodically made available at https://dumps.wikimedia.org/trwiki/20190120/.

In addition to BBC and DW, the following international media outlets produce content in Turkish: Global Voices (https://globalvoices.org/), The Voice of America (https://www.amerikaninsesi.com/) , China Plus (formerly China Radio International, http://turkish.cri.cn/), and TWR360 (https://www.twr360.org/, although most of it is multimedia content). These outlets may be interesting sources from which to obtain monolingual corpora.

Bilingual corpora: OPUS[25] contains ready-made corpora for Turkish (see table 13) and reports a total 151.9 million parallel sentences. Most of it comes from OpenSubtitles. It is unclear whether OpenSubtitles and Global Voices versions are incremental; in this case, the total would be reduced to around 50 milllion words.

[25] http://opus.nlpl.edu

corpus	doc's	sent's	en tokens	tr tokens
OpenSubtitles v2018	58,210	47.4M	374.3M	274.7M
OpenSubtitles v2016	47,266	39.6M	312.3M	228.5M
OpenSubtitles v2012	25,455	22.6M	178.2M	129.6M
OpenSubtitles v2013	22,631	20.6M	164.1M	119.0M
OpenSubtitles v2011	21,091	18.4M	146.3M	106.5M
Tanzil (Quran) v1	135	1.3M	25.4M	20.8M
SETIMES v2	1	0.2M	5.1M	4.5M
Wikipedia v1.0	2	0.2M	4.8M	4.2M
SETIMES v1	1	0.2M	4.5M	4.2M
TED2013 v1.1	1	0.1M	2.7M	2.0M
GNOME v1	1,235	0.5M	2.3M	2.3M
Tatoeba v2	1	0.2M	3.6M	1.0M
OpenSubtitles v1	197	0.2M	1.9M	1.3M
KDE4 v2	1,285	0.2M	1.1M	0.8M
EUbookshop v2	67	24.1k	0.9M	0.7M
Ubuntu v14.10	459	0.1M	0.8M	0.5M
PHP v1	2,855	38.5k	0.5M	0.1M
GlobalVoices v2017q3	192	5.6k	0.1M	96.7k
GlobalVoices v2015	149	4.0k	90.3k	70.7k
total	181,233	151.9M	1.2G	900.8M

Table 13: OPUS resources for Turkish–English.

2.15.4 Resources

Monolingual resources: Apertium has fair-coverage morphological resources for Turkish (https://github.com/apertium/apertium-tur: 17,721 lemmata, 76 disambiguation rules, 92.2% coverage on the SETimes corpus[26], 82.3% on Wikipedia dumps (2013)).

There is a GPL-licensed morphological analyser for Turkish, TRmorph (http://wiki.apertium.org/wiki/Trmorph) with about 37,300 stems and 90% coverage on SETimes.

The Unimorph project (https://github.com/unimorph) contains a set of morphologically-annotated Turkish forms (https://github.com/unimorph/tur); the coverage of the 2018 version of the Turkish news crawls described above is low (14%) with 213,540 forms.

Crawls of Turkish news text are available at http://data.statmt.org/news-crawl/tr/.

Bilingual resources: A crowd-sourced Turkish–English dictionary (http://www.seas.upenn.edu/~nlp/resources/TACL-data-release/dictionaries.tar.gz) has been made available by Pavlick et al. (2014).

The following online commercial MT systems offer Turkish–English:

- Bing Translator (https://www.bing.com/translator)
- Google Translate (https://translate.google.com)
- PROMT Online (https://www.online-translator.com)
- Yandex Translate (https://translate.yandex.com)

2.15.5 Challenges for corpus-based MT to English

The main challenges when translating from English to Turkish come from grammatical differences:

- agglutination in noun-based phrases, especially in the form of case or possessive suffixes;
- very complex verb forms;
- generation of affixes for features not explicitly encoded in English such as perfective/imperfective aspect, evidentiality, cases, etc.
- absence of gender marks in 3rd-person pronouns;
- radically different sentence and phrase structures —position of object, obliques, and verb, use of postpositions, relative clauses and genitives before noun, etc.—;

Turkish is a highly inflected language. This can cause data sparseness problems, but mainly if the MT system treats the words as atomic units. It may therefore be desirable that that the different grammatical suffixes are represented as independent tokens to allow the system to generalize better from the training data, or that an unsupervised sub-word strategy is learned. Moreover, the absence of specific news-related bilingual corpora may be an obstacle to good results in a content creation task.

[26] http://nlp.ffzg.hr/resources/corpora/setimes/

2.16 Yoruba (yo, yor)

2.16.1 Factsheet

According to Wikipedia, Yoruba has around 28 million speakers. It is spoken (and official) in Nigeria, Benin, and Congo. It belongs to a very large family, the Niger–Congo family, as Swahili (§ 2.13), and more specifically to the Volta–Niger group, as Igbo (§ 2.7).

Yoruba is written in a variety of writing systems, all based on the Latin alphabet. As Yoruba vowels may be uttered in three different *tones*, high, middle and low. A change in tone may mean a change in meaning (*bàtà*, low–low, means 'shoe'; *bàtá*, low–high, refers to a kind of drum). Spelling systems may or may not mark tones with diacritics, leading to lexical ambiguity; examples in this deliverable may also be inconsistent in this respect. Reputed sources such as the BBC news in Yoruba site (https://www.bbc.com/yoruba) sometimes decide not to mark tone in some part of a news piece: one can find, for instance *Nàìjíríà* ('Nigeria') in headlines or in the first few paragraphs of a news piece along with *Naijiria* in the remaining text. BBC acknowledges that they are aware that the official way to write Yoruba is with all tone marks included, and that is done for as much as possible of the headline and first three paragraphs, as many readers do not read any further, and that the decision to mark tones or not may also depend on the target audience of each particular, as younger people feel more comfortable with Yoruba without tone marks.

2.16.2 Contrasts with English

Nouns			
Feature	**Value in English**	**Value in Yoruba**	**Examples**
Coding of plurality in nouns	Plural suffix	Plural word	*ilé* ('house') vs. *àwọn ilé* ('houses', i.e. 'PLURAL-MARKER house')
Occurence of plurality in nouns	All nouns, always obligatory	All nouns, always optional	*ilé púpọ̀* ('many houses', lit. 'house many') vs. *àwọn ilé* ('houses', i.e. 'PLURAL-MARKER house')

Gender			
Feature	**Value in English**	**Value in Yoruba**	**Examples**
Number of genders	Three (sex, based, only in singular pronouns and possessives)	None	English has *she*, *he*, and *it*, where Yoruba uses only *ó*.

Verbs			
Feature	**Value in English**	**Value in Yoruba**	**Examples**
Person marking in verbs	Only the A (≃agent) argument (3rd person singular, to be)	No person marking	*Mo kọrin* ('I sing', 'I sang'); *O kọrin* ('He sings', 'He sang')
Passive constructions	Present	Absent	Yoruba would use constructions with *a* ('they') as subject when no agent is provided: *A ri mi* ('I am seen', lit. 'They see me')
Verb inflection	Some inflection with person, number and tense.	Invariable verbs (but contractions with some particles o	In Yoruba, particles occur between the subject and the verb. The bare verb usually indicates a past, completed action. Examples: *ń*, imperfective/progressive; *ti*, perfective; *á*, future. Some of these particles may combine: *mo ti ń gba lẹ́tà rẹ* 'I have started to receive your letters' (lit. I 'PERFECTIVE IMPERFECTIVE receive letter your')[27]
Nominal and locational predication ('to be something' vs. 'to be somewhere')	Identical	Different	English uses the same verb for nominal predication ('She is an Engineer') and locational predication ('She is in the restaurant'). Yoruba has different verbs.

Function words			
Feature	**Value in English**	**Value in Yoruba**	**Examples**
Definite and indefinite articles	Definite article (word distinct from demonstrative) and indefinite article (word distinct from 'one')	No definite or indefinite articles	*ilé* may be 'house', 'the house' or 'a house'.

[27] http://www.languagesgulper.com/eng/Yoruba.html

Pronouns			
Feature	**Value in English**	**Value in Yoruba**	**Examples**
Gender distinctions in independent personal pronouns	3rd person singular only	No definite or indefinite	*Ó ri mí* may be 'he saw me' or 'she saw me'.
Politeness distinction in pronouns	No distinction	Binary politeness distinction	

Word order and syntax			
Feature	**Value in English**	**Value in Yoruba**	**Examples**
Order of genitive and noun	Both occur	Noun–Genitive	"fìlà Àkàndé" ('Akande's cap'), "Àáre Nàìjíríà" ('[the] president of Nigeria')
Order of adjective and noun	Adjective–noun	Noun-adjective	*ilé ńlá* ('large house', lit. 'house large')
Order of demonstrative and noun	Demonstrative–noun	Noun–demonstrative	*ilé naa* ('that house', lit. 'house that')
Order of numeral and noun	Demonstrative–numeral	Numeral–demonstrative	*ilé naa* ('that house', lit. 'house that')
Polar questions	Interrogative word order differs from affirmative word order	Uses intonation for word order	*O ri mi* ('You see me') vs. *Ṣe o ri mi?* ('Do you see me')
Verb chains or verb series	Very uncommon	Very common	*Ó jẹun sùn* 'He ate before going to sleep' (lit. 'He ate slept'); *Ó jókòó mu ọti* ('He sat to drink a glass', lit. 'He sat drank a glass') (Sachnine and Akinyemi, 1997, 30).

Contractions: This could be a potential issue with some Yoruba text, although its extent is not clear. In writing, adjacent words contract into a single word, usually with a loss of vowels (elision) and other phonological changes. Verbs contract with objects as in *Ó fọṣọ* from *Ó fọ aṣọ* 'He/she washes the clothes', sometimes with phonological alternations as *Ó lówó* from *Ó ní owó*[28] ('He/she has money'). This would increase the effective vocabulary considerably, as contractions

[28]The sounds *n* and *l* alternate in nasal and non-nasal contexts.

affect content words in open classes. However, neither Google Translate nor Glosbe (see 'Bilingual resources' below) show these contractions, which are however described by Sachnine and Akinyemi (1997, 24).

2.16.3 Corpora

Monolingual Corpora: The Yoruba Wikipedia[29] has about 32,000 articles containing about 1.2 million words.[30] The monolingual dump of the Yoruba wikipedia is periodically made available at https://dumps.wikimedia.org/yowiki/.

In addition to the BBC (https://www.bbc.com/yoruba), we have only identified Global Voices (https://yo.globalvoices.org/) as an international media outlet producing content in Yoruba.

Sketch Engine announces it has a YorubaWaC corpus (crawled with SpiderLing and WebBootCat) with 2.8 million words, but it is not available and licensing is not clear.[31]

Bilingual corpora: Publicly available corpora are scarce. OPUS[32] has only about 25,000 domain-specific sentence pairs from GNOME and Ubuntu.

2.16.4 Resources

Monolingual resources: A free/open-source morphology for Yoruba is announced as being available from the first author of an EACL-2009 paper (Finkel and Odejobi, 2009) upon request.

Bilingual resources: A crowd-sourced Yoruba–English dictionary (http://www.seas.upenn.edu/~nlp/resources/TACL-data-release/dictionaries.tar.gz) has been made available by Pavlick et al. (2014).

Google Translate has a Yoruba translator. It generally generates Yoruba without diacritics, or it does so inconsistently (for example, translations for the word 'snake' are sometimes *ejò* and some other times *ejo*. Also, it does not perform verb–object contractions such as the ones described above.

Glosbe has an interface to search Yoruba–English dictionaries and examples (https://glosbe.com/yo/en, https://glosbe.com/en/yo).

2.16.5 Challenges for corpus-based MT from English

One could say that Yoruba is not too different from English. Both are subject–verb–object languages which are not too inflected. When translating from English to Yoruba, some of the contrasts described above in 2.16.2 are not relevant (for instance, the three genders of English 3rd-person pronouns and possessives, or the absence of definite and indefinite articles. In summary:

[29] https://yo.wikipedia.org/
[30] https://yo.wikipedia.org/wiki/Special:Statistics?action=raw
[31] https://www.sketchengine.eu/yowac-yoruba-corpus/
[32] http://opus.nlpl.eu

- There is a sheer scarcity of parallel corpora, and it remains to be seen how much more bilingual text may be crawled using tools such as Bitextor.[33] Approaches based on monolingual corpora (Artetxe et al., 2018) may be unfeasible unless the project manages to crawl Yoruba (for instance, using Spiderling(Suchomel et al., 2012)) texts beyond a few million words. Another approach worth trying would be synthetic bilingual corpora (perhaps back-translated using commercial systems such as Google translate).
- Word-order differences seem to occur locally (basically inside the noun phrase). This may only be a problem for longer noun phrases.
- The Yoruba side of corpora may be inconsistent as regards tone diacritics. This may be a problem when generating Yoruba. One possible solution would be to strip all diacritics at train time and then use a corpus-trained standalone diacriticizer.
- Getting the right form of Yoruba verbs (beyond simple present and past) can be sometimes tricky. Categories such as aspect (finished versus unfinished actions) are more important in Yoruba (where they are marked with auxiliaries) than tense (past versus present), and mappings are asymmetrical. Also,
- Yoruba does not have plural forms. It uses an auxiliary word (àwọn) but only where the plural meaning cannot be inferred from other words in the sentence or context. English plural nouns may have to be mapped to Yoruba structures where plural is not marked.
- Yoruba does not have a passive while English uses it often; it has to be mapped to impersonal *they* (Yoruba *a*) when there is not an agent, and converted to some active structure when there is an agent.
- English interrogatives have to be reordered when translating to Yoruba.
- Politeness distinctions in 2nd-person pronouns will be hard to generate in Yoruba as English does not mark politeness.
- Selecting a different translation for the verb ‘to be’ for ‘to be something’ and for ‘to be somewhere’ may pose some problems as both usages are quite abundant.
- In some texts, verbs may be written contracted with their objects, with elision phenomena and even some consonant changes, leading to a sharp increase in effective vocabulary size unless a sub-word approach is used.

[33] https://github.com/bitextor/bitextor/

3 Conclusion

This deliverable has described the available monolingual and bilingual resources and corpora, and the challenges that project GoURMET may expect to face when building neural machine translation between English and the sixteen low-resource languages of interest to the project GoURMET. Each description includes a quick fact sheet for the language, as well as a linguistic description of the main contrasts between the language and English.

The systematic description of the languages in this deliverable and the resources available for them will inform discussions about the languages to cover in years two and three of the project, as well as to document for researchers the challenges and opportunities available for these language pairs.

References

Akinremi, I. I. (2013). Impersonal constructions in igbo. *Theory and Practice in Language Studies*, 3(7):1129.

Anagbogu, P. N. (1995). The semantics of reduplication in igbo. *Journal of West African Languages*, 25(1):43–52.

Artetxe, M., Labaka, G., and Agirre, E. (2018). Unsupervised statistical machine translation. *arXiv preprint arXiv:1809.01272*.

Ataman, D. (2018). Bianet: A parallel news corpus in Turkish, Kurdish and English. In *LREC Workshop MLP-MomenT 2018*.

Birch, A., Osborne, M., and Koehn, P. (2008). Predicting success in machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 745–754. Association for Computational Linguistics.

Campbell, G. L. and King, G. (2010). *The Routledge concise compendium of the world's languages*. Routledge.

De Pauw, G., Wagacha, P. W., and De Schryver, G.-M. (2011). Exploring the sawa corpus: collection and deployment of a parallel corpus english—swahili. *Language resources and evaluation*, 45(3):331.

Esmaili, K. S. (2012). Challenges in Kurdish text processing. *CoRR*.

Esplà-Gomis, M., Klubička, F., Ljubešić, N., Ortiz-Rojas, S., Papavassiliou, V., and Prokopidis, P. (2014). Comparing two acquisition systems for automatically building an english—croatian parallel corpus from multilingual websites. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1252–1258, Reykjavik, Iceland. European Language Resources Association (ELRA).

Finkel, R. and Odejobi, O. A. (2009). A computational approach to yorubá morphology. In *Proceedings of the First Workshop on Language Technologies for African Languages*, pages 25–31. Association for Computational Linguistics.

Gökırmak, M. and Tyers, F. M. (2017). A dependency treebank for Kurmanji Kurdish. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 64–72.

Hassani, H. (2018). BLARK for multi-dialect languages: Towards the Kurdish BLARK. *Lang. Resour. Eval.*, 52(2):625–644.

Jaggar, P. J. (2001). *Hausa*, volume 7. John Benjamins Publishing.

Jindal, S., Goyal, V., and Bhullar, J. S. (2017). Building english-punjabi parallel corpus for machine translation. *International Journal of Computer Applications*, 180(8):26–29.

Klubička, F., Toral, A., and Sánchez-Cartagena, V. M. (2018). Quantitative fine-grained human evaluation of machine translation systems: a case study on english to croatian. *Machine Translation*, 32(3):195–215.

Lipps, J. (2011). xsma: A finite-state morphological analyzer for swahili.

Ljubešić, N., Fišer, D., and Erjavec, T. (2014). TweetCaT: a Tool for Building Twitter Corpora of Smaller Languages. In Chair), N. C. C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Ljubešić, N. and Klubička, F. (2014). {bs,hr,sr}WaC – web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35, Gothenburg, Sweden. Association for Computational Linguistics.

Ljubešić, N. and Erjavec, T. (2016). Corpus vs. lexicon supervision in morphosyntactic tagging: the case of slovene. In Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Ljubešić, N., Klubička, F., Željko Agić, and Jazbec, I.-P. (2016). New inflectional lexicons and training corpora for improved morphosyntactic annotation of croatian and serbian. In Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Marinov, S. and Nivre, J. (2005). A data-driven dependency parser for bulgarian. In *Proc. of the 4th Workshop on Treebanks and Linguistic Theories (TLT)*, pages 89–100.

Park, J., Hong, J.-P., and Cha, J.-W. (2016). Korean Language Resources for Everyone. In *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation (PACLIC 30)*, pages 49–58, Seoul, Korea.

Pavlick, E., Post, M., Irvine, A., Kachaev, D., and Callison-Burch, C. (2014). The language demographics of amazon mechanical turk. *Transactions of the Association for Computational Linguistics*, 2:79–92.

Perrott, D. V. (1965). *Teach Yourself Swahili*. english universities Press.

Sachnine, M. and Akinyemi, A. (1997). *Dictionnaire yorùbá-français: suivi d'un index français-yorùbá*. KARTHALA Editions.

Saveski, M. and Trajkovski, I. (2010). Development of an English-Macedonian machine readable dictionary by using parallel corpora. In *International Conference on ICT Innovations 2010*, pages 195–204.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1715–1725.

Simov, K., Osenova, P., Slavcheva, M., Kolkovska, S., Balabanova, E., Doikoff, D., Ivanova, K., Simov, A., and Kouylekov, M. (2002). Building a linguistically interpreted corpus of bulgarian: the bultreebank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Song, K. A. (2010). Various evidentials in Korean. In *Proceedings of the 24th Pacific-Asia Conference on Language, Information and Computation*.

Suchomel, V., Pomikálek, J., et al. (2012). Efficient web crawling for large text corpora. In *Proceedings of the seventh Web as Corpus Workshop (WAC7)*, pages 39–43.

Tedla, Y. K., Yamamoto, K., and Marasinghe, A. (2016). Nagaoka tigrinya corpus: Design and development of part-of-speech tagged corpus. In *Proceedings of the 22nd Annual Conference of the Speech Processing Society of Japan*, pages 413–416.

Tesfaye, D. and Abebe, E. (2010). Designing a rule based stemmer for afaan oromo text. *International journal of computational linguistics (IJCL)*, 1(2).

Thackston, W. M. (2006a). *Kurmanji Kurdish:-A Reference Grammar with Selected Readings*. Renas Media.

Thackston, W. M. (2006?b). Sorani kurdish:-a reference grammar with selected readings.

Tyers, F. and Serdar Alperen, M. (2010). South-East European Times: A parallel corpus of Balkan languages. In *Proceedings of the Workshop on Exploitation of multilingual resources and tools for Central and (South) Eastern European Languages, LREC 2010*.

Ugochukwu, F. (2004). *Dictionnaire igbo-français: suivi d'un index français-igbo*. KARTHALA Editions.

Veisi, H., MohammadAmini, M., and Hosseini, H. (2019). Toward Kurdish language processing: Experiments in collecting and processing the AsoSoft text corpus. *Digital Scholarship in the Humanities*.

Virk, S., Humayoun, M., and Ranta, A. (2011). An open-source punjabi resource grammar. In *Proceedings of RANLP-2011, Recent Advances in Natural Language Processing, Hissar, Bulgaria, 12-14 September, 2011*, pages 70–76.

Walther, G. and Sagot, B. (2010). Developing a large-scale lexicon for a less-resourced language: General methodology and preliminary experiments on Sorani Kurdish. In *SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages (LREC 2010 workshop)*.

Walther, G., Sagot, B., and Fort, K. (2010). Fast development of basic NLP tools: Towards a lexicon and a pos tagger for Kurmanji Kurdish. In *Proceedings of the 29th International Conference on Lexis and Grammar*.

Wegari, G. M. (2011). Parts of speech tagging for afaan oromo. *International Journal of Advanced Computer Science and Applications Special Issue on Artificial Intelligence*.

ENDPAGE

GoURMET

H2020-ICT-2018-2 825299

D1.1 Survey of relevant low-resource languages