Comparable Corpora: Opportunities for New Research Directions

Kenneth Church
Northeastern University
k.church@northeastern.edu

Abstract

Most conference papers present new results, but this paper will focus more on opportunities for the audience to make their own contributions. This paper is intended to challenge the community to think more broadly about what we can do with comparable corpora. We will start with a review of the history, and then suggest new directions for future research.

1 Introduction

The success of chat bots in many languages demonstrates the power of comparable corpora (CC) and pivoting via English. We will start with a review of the history of CC, and then suggest new directions for future research:

1. More depth: CC are normally used for simple tasks such as bilingual lexicon induction (BLI), but CC can be used for much more interesting views of lexical semantics.
2. Compare and Contrast: CC are normally used to make simple comparisons over language pairs, but they can be used for contrasts as well as comparisons (in monolingual settings as well as multilingual settings).
3. More modalities: Now that vectors encode everything (text in many languages, pictures, audio, video), we can compare and contrast everything with everything.
4. Bursting filter bubbles: bots made in America are trained on corpora from an American perspective with American biases. We should not impose American values on others.

2 Historical Background

2.1 Parallel Corpora

Table 1 shows some examples of parallel corpora. HF and LDC in Table 1 refer to HuggingFace[1] and the Linguistic Data Consortium,[2] respectively. More parallel corpora can be found on HF by searching for *parallel*, *aligned* and *translation*.

The main application of parallel corpora has been machine translation (Brown et al., 1993). Shannon's noisy channel model (Shannon, 1948) was originally motivated for applications in communication (telephones), but it has been used for many other applications including machine translation. That is, to translate from English to French, one imagines that French speakers think in English, like English speakers do, but for some reason, when French speakers talk, the noisy channel converts their English to French.

$$E \rightarrow Noisy\ Channel \rightarrow F \tag{1}$$

The task of the translation system is to recover the original English, E, from the observed French, F. These days, it has become standard practice to use

[1] https://huggingface.co/

[2] https://www.ldc.upenn.edu/

Resource	Cites
Europarl (Koehn, 2005)	4634
OPUS (Tiedemann, 2012)	2255
HF: Helsinki-NLP/opus-100	
(Resnik and Smith, 2003)	848
MultiUN (Eisele and Chen, 2010)	327
HF: Helsinki-NLP/multiun	
Bible (Pratap et al., 2024)	254
(Akerman et al., 2023)	
HF: Flux9665/BibleMMS	
LDC: Hansard French/English	
LDC: Hong Kong Hansards	
HF: NilanE/ParallelFiction-Ja_En-100k	
HF: sentence-transformers/parallel-sentences	
HF: tiagoblima/bible-ptbr-gun-gub-aligned	
HF: dsfsi/vukuzenzele-sentence-aligned	
(Marivate et al., 2023)	

Table 1: Examples of parallel corpora

Proceedings of the 18th Building and Using Comparable Corpora Workshop, pages 73–82
January 20, 2025.

English	French	Sense
bank	banque	money
	banc	river
duty	droit	tax
	devoir	obligation
drug	médicament	medical
	drogue	illicit
land	terre	property
	pays	country
language	langue	medium
	langage	style
position	position	place
	poste	job
sentence	peine	judicial
	phrase	grammatical

Table 2: Using Hansards for Word Sense Disambiguation (WSD), based on Table 2 in Gale et al. (1992)

neural networks for translation, but it used to be popular to use Hidden Markov Models (HMMs) to find the most likely English, $\hat{E}$, based on a prior (language model), $Pr(E)$, and a bilingual dictionary, $Pr(F|E)$.

$$\hat{E} = \text{argmax}_E Pr(E)Pr(F|E) \qquad (2)$$

Much of the discussion below will focus on the bilingual lexicon, $Pr(F|E)$. $Pr(E)$, the language model in Eqn (2), is relatively well estimated because we can re-use monolingual LLMs that have been developed for other applications. The bilingual lexicon, $Pr(F|E)$, assigns probabilities to all sequences of English, E, and French, F. It was standard practice, at least at first, to estimate $Pr(F|E)$ from parallel corpora such as the English-French Canadian Hansards.

The rest of this section on history will largely focus on the lexicon. After introducing comparable corpora as an alternative to parallel corpora, we will motivate WSD (word-sense disambiguation). Much of the research on WSD started with bilingual word-senses, but it should be noted that word-senses are different in monolingual and bilingual dictionaries.

There has also been considerable work on transferring monolingual lexical resources such as WordNet and VAD to more languages. Unfortunately, much of this work uses translation to pivot out of English in inappropriate ways, as we will see.

This section will end with a review of BLI (bilingual lexicon induction). The BLI literature uses more modern methods in machine learning than previous methods for inducing lexicons from CC, but BLI benchmarks (such as MUSE) may not be as effective as older WSD methods for addressing classic challenges with translations of ambiguous words. A classic example is *bank*, which is translated as *banque* and *banc* in the Canadian Hansards, depending on the sense. Unfortunately, this ambiguity is not captured in the MUSE benchmark where *bank* translates to *banque* (but not *banc*). In the reverse direction, MUSE has translations for both *banque* and *banc*, but they translate to different English words, *bank* and *bench*, respectively. Comparisons of ambiguities in Hansards (Table 2) and MUSE (Table 5 and Table 6) suggest that MUSE is not testing WSD as much as the older literature. Another concern with MUSE is that most words in the benchmark translate to themselves. These concerns suggest that there may be room to introduce a new benchmark that would make a stronger case for comparable corpora (CC).

After discussing history, the next section will discuss more radical challenges for the future: lexical semantics, transfer learning, filter bubbles and connections between academic search and CC.

2.2 Comparable Corpora (CC)

The term, *comparable corpora*, was introduced in (Fung and Church, 1994; Rapp, 1995; Fung and Yee, 1998; Fung, 2000) to address limitations with parallel corpora. Parallel corpora are available for a few genres such as parliamentary debates (Hansards) and religion (Bible), as shown in Table 1. Since most texts and most genres are not translated, we can collect larger and more diverse corpora if we relax the restriction on translation. CC replace a single parallel corpus with two monolingual corpora, ideally on similar (comparable) topics.

2.3 Word-Sense Disambiguation (WSD)

In addition to machine translation applications mentioned above, parallel corpora have also been used to disambiguate ambiguous words such as *bank*, as illustrated in Table 2. Bar-Hillel (1960) thought machine translation was impossible when he could not figure out how to disambiguate words such as those in Table 2. It was obvious that the translation depends on a solution to WSD.

Gale et al. (1992) used this argument in reverse to obtain large quantities of labeled text for WSD research. They used parallel corpora such as Hansards to find instances of ambiguous words such as *bank*, and use the French translations to label each instance of *bank* as either "money" sense or "river" sense. After labeling the English in this way, they threw away the French and used the sense-labeled text to train and test machine learning methods for WSD.

2.4 Monolingual Senses != Bilingual Senses

This approach was successful in reviving interest in WSD research, though it should be mentioned that bilingual lexicography is different from monolingual lexicography. Consider the word *interest*. This word has many senses including a "money" sense and a "love" sense, among others. A monolingual dictionary will describe each of these senses in considerable detail. However, there will be little to say about *interest* in an English-French bilingual dictionary because the same complications are shared between the English word and its French equivalent. Thus, the approach above is more effective for words like those in Table 2 where the word is ambiguous in one language but not the other, and less effective for words like *interest*, which are equally ambiguous in both languages.

2.5 Inappropriate Uses of Translation

Parallel corpora are limited in a number of ways. Genre is perhaps the most obvious limitation, but a more serious limitation may be distortions introduced by translation.

When I was first working with Hansards in the 1990s, I tried to pitch parallel corpora to Sue Atkins, a lexicographer who specialized in English-French bilingual dictionaries. She rejected my pitch, objecting to "translationese"[3] as "unnatural" natural language. In addition, she criticized concordance tools for parallel corpora because they failed to distinguish source and target languages. Examples of these tools can be found in the sketch engine;[4] these tools show examples of a word in one language as well as its equivalents in other languages.

2.5.1 XNLI: A Multilingual version of NLI

Much of the work on parallel corpora treats the source and target languages as equivalent (with equal status), ignoring distortions introduced by translation. We should be more careful about translation artifacts in many benchmarks. Artetxe et al. (2020) call out XNLI, a English version of an NLI task. The monolingual NLI task depends on word overlaps between the premise and the hypothesis, but many of these crucial overlaps are lost in translation in the XNLI version where premises and hypotheses are translated independently. Too much of the work in computational linguistics uses translation to pivot via English in inappropriate ways.

Synset	French Glosses
dog.n.01	canis_familiaris, chien
cat.n.01	chat
house.n.01	maison
bank.n.01	banque, rive

Table 3: Global WordNet pivots from English

English	Hausa	V	A	D
aaaaaaah	aaaaaaa	0.48	0.61	0.29
aaaah	aaaah	0.52	0.64	0.28
aardvark	ardvark	0.43	0.49	0.44
aback	abin mamaki	0.39	0.41	0.29
abacus	abacus	0.51	0.28	0.49
abalone	abalone	0.50	0.48	0.41

Table 4: NRC-VAD pivots from English using Google Translate; V = Valance, A = Arousal & D = Dominance

2.5.2 No Language Left Behind (NLLB)

Abdulmumin et al. (2024) report serious problems with FLORES (Goyal et al., 2022) in four African languages. A common problem was the use of Google Translate, which sometimes produced "incoherent or unclear" Hausa text. FLORES is an important test set for NLLB (no language left behind) (NLLB Team et al., 2022).

2.5.3 WordNet and VAD

Table 3 and Table 4 show attempts to use translation to pivot from English to other languages. WordNet[5] (Miller, 1995) and NRC-VAD (Mohammad, 2018)[6] were originally designed for English. Translation was used to transfer them to more languages. Note that translation introduces losses; bank.n.01 cannot be both the money sense (*banque*) and the river sense (*rive*). I asked a colleague, a native speaker of Hausa, to comment on Table 4. None of the Hausa words in the table are that useful. Most of the words in the Hausa column are English, with the exception of *abin mamaki* which Google translates

[3] https://en.wiktionary.org/wiki/translatese

[4] https://www.sketchengine.eu/guide/parallel-concordance-searching-translations/

[5] https://www.nltk.org/howto/wordnet.html

[6] https://saifmohammad.com/WebPages/nrc-vad.html

to *what a surprise* in English. My informant did not know what *aback* means in English even though his English is excellent. When I explained it to him, we agreed that this translation is not convincing.

English	French
bank	banque, banques, *but not* banc
duty	devoir, *but not* droit
drug	drogue, médicament
land	terre, terrain, terres, *but not* pays
language	langue, langues, langage
position	position, *but not* post
sentence	peine, phrase, sentence
good	bien, bon, bonne, bonnes, bon
bad	mal, mauvais, mauvaise, bad

Table 5: Some examples from MUSE: fr $\rightarrow$ en

In short, there are many problems with using translation to pivot from English to many other languages. It is unlikely that the structure of the English WordNet ontology and the English VAD lexicon is universal over all languages. In the West, we slay dragons, but in the East, dragons are good luck. In the West, white is common for weddings and black is common for funerals, but in some places, white is common for funerals, and in other places, red is common for weddings. Even the list of concepts is likely to vary from one language to another. Many of the English words in Table 4 are not (much of) "a thing" in Hausa.

French	English
banc	bench, *but not* bank
banque	bank, banking
droit	right, law
devoir	duty
drogue	drug, drugs, drogue
médicament	medicine, drug, medication
terre	land, earth, soil, terre
terrain	land, terrain
terres	land, lands
langue	language
langues	language, languages
langage	language
position	position
peine	sentence, pain, penalty, sorrow
phrase	sentence, phrase
sentence	sentence, sentencing

Table 6: Some examples from MUSE: en $\rightarrow$ fr

Dict	Pairs	Src	Tgt	Src=Tgt
en $\rightarrow$ fr	113,286	94,681	97,035	73,471
fr $\rightarrow$ en	113,324	97,021	94,730	73,471

Table 7: MUSE Dictionary Sizes

2.6 Bilingual Lexicon Induction (BLI)

Much of the work on BLI is based on the MUSE benchmark [7] (Lample et al., 2017; Conneau et al., 2017). The MUSE benchmark provides:

1. fastText[8] embeddings for 30 languages, and
2. gold set of bilingual dictionaries, $D_{l_i \rightarrow l_j}$ for 110 pairs of languages: l_i, l_j. The gold sets are split into training (seed) dictionaries and test dictionaries.

See section 2.2 of (Sharoff et al., 2023) for an introduction to vector space models and CC. The fastText embeddings, $X_l \in \mathbb{R}^{|V_l| \times d}$, contain a row for each word in the vocabulary, V_l, for language l. The rows are vectors of length d, where d is the number of hidden dimensions.

Each dictionary, $D_{l_i \rightarrow l_j}$ consist of a list of pairs of words in the two languages. Table 7 counts the number of pairs in both directions, as well as the number of unique words in the source language (src) and target language (tgt). Many of the pairs use the same word in both languages, as indicated by the last column.

The task is to estimate a dictionary, $\hat{D}_{l_i \rightarrow l_j}$, for a pair of languages, l_i and l_j. We then compare estimates, $\hat{D}$, with gold dictionaries, D. A simple approach is to use the training (seed) dictionaries to estimate a rotation matrix, $R \in \mathbb{R}^{d \times d}$, where $R = \text{argmin}_R \, ||RX_{l_i} - X_{l_j}||_F^2$. It is standard practice to estimate R with the orthogonal Procrustes problem[9] (Schönemann, 1966).

At inference time, we start with a vector in X_{l_i}, and then rotate those vectors by R and use approximate nearest neighbors (ANN) (Bruch, 2024) to find nearby vectors in X_{l_j}.

Early work on CCs attempted to collect word lists (Kilgarriff et al., 2014) and infer bilingual lexicons; MUSE updates this approach using modern methods in machine learning. That said, MUSE may not be as effective as older methods for WSD because of gaps. Examples from $D_{\text{fr}\rightarrow\text{en}}$ and

[7] https://github.com/facebookresearch/MUSE
[8] https://github.com/facebookresearch/fastText
[9] https://docs.scipy.org/doc/scipy/reference/generated/scipy.linalg.orthogonal_procrustes.html

$D_{\text{en}\rightarrow\text{fr}}$ are shown in Tables 5-6; some of the ambiguities in Table 2 are covered, and some are not. An example of a gap is: *bank* (en) $\rightarrow$ *banc* (fr); this pair is missing from both $D_{\text{en}\rightarrow\text{fr}}$ and $D_{\text{fr}\rightarrow\text{en}}$.

3 Challenges for the Future

3.1 BLI, PMI and Lexical Semantics

Much of the work on BLI uses a simple view of a bilingual lexicon where single words in one language correspond to single words in another language, more or less one-for-one. Obviously, the relationship is far more complicated than this. The phrasal verb, *ask for*, is similar to *request*, violating the one word for one word assumption.

3.1.1 Etymology

More seriously, there is a difference in register, going back to the Norman Conquest in 1066. For a few hundred years after 1066, the English Court spoke French. As a result, English borrowed many words from French. The French term typically has a higher register than the older English equivalent; the peasants raise *cows, calf* and *swine* so the aristocracy can eat *beef, veal* and *pork*.[10]

3.1.2 Distributional Methods

Much of the work on BLI does not take advantage of etymology because work on BLI is based on the Distributional Hypothesis[11] (Harris, 1964) and Firth's "You shall know a word by the company it keeps" (Firth, 1957). The distributional hypothesis is convenient for computation, suggesting "(unlabeled) corpora are all we need," though many aspects of linguistics go beyond distributional evidence, e.g., etymology, lexical semantics.

There are interesting connections between popular distributional methods, e.g., PMI (pointwise mutual information), Word2Vec and LLMs (large language models). The connection between BLI and Word2Vec was mentioned above. Levy and Goldberg (2014) view Word2Vec as a factored representation of PMI (Church and Hanks, 1990). BERT and chat bots can be viewed as an enhancement of Word2Vec; instead of representing words as vectors, we now represent sequences of 512-subword units as vectors.

[10] https://www.csmonitor.com/The-Culture/In-a-Word/2021/0510/They-re-cows-in-the-field-but-beef-on-the-table

[11] https://aclweb.org/aclwiki/Distributional_Hypothesis

Relation	PMI	Lexical Sem	Back Trans
Synonyms	large	$=$ (equiv. rel.)	large
Antonyms	large	$\neq$ (anti-sym)	small
Is-a	small	$\leq$ (partial order)	small
Part-Whole	large		small

Table 8: PMI $\neq$ Lexical Semantics

3.1.3 Lexical Semantics

As mentioned above, lexical semantics is a challenge for distributional methods. While there are some similarities between PMI (collocations) and lexical semantics (synonyms, antonyms, is-a), there are also some important differences, as shown in Table 8. PMI scores are large when words appear near one another more than chance. Consequently, both synonyms and antonyms have large PMI scores because documents often compare and contrast this with that. Similarly, PMI scores can be large for other words that appear near one another, e.g., *window, door* and *house*. Large PMI scores do not necessarily imply synonymy.

Back translations are also mentioned in Table 8. Back translations are more effective than PMI for distinguishing synonyms from antonyms. If we take a random walk over MUSE dictionaries and start from *good*, such walks will often take us to synonyms, but rarely to antonyms. There is an opportunity to propose a theory of translation and collocation based on linear algebra and graph theory. This theory should explain the observations in Table 8 where antonyms are close in terms of PMI but not in terms of random walks on translations.

3.1.4 Avoid Pivoting via English

As mentioned in subsection 2.4, monolingual lexicography is different from bilingual lexicography. For example, *interest* has many senses in monolingual dictionaries, but not in bilingual dictionaries. *Bank* is ambiguous in English, but not in French. Bilingual dictionaries become interesting when the senses are not isomorphic. Table 3 and Table 4 take an overly simplistic view of the structure of the lexicon where the ontology (and VAD values) are assumed to be universal. Translating from English is likely to introduce distortions. Can we do better than pivoting via English?

3.2 Transfer Learning

Suppose we want to transfer from a high resource language such as English to growth opportunities such as Indonesian (id) and Hausa (ha). We prefer

Language	Wikipedia	Joshi	S2 Abstracts	ACL	HF Datasets	HF Models	Speakers
en	6,917,939	5	88,348,938	103,000	10,749	50,717	1456M
zh	1,452,669	5	3,061,847	71,800	1202	4495	1138M
hi	163,524	4	2,848	8,740	421	1388	610M
es	1,992,685	5	2,742,468	28,600	945	3245	559M
fr	2,650,236	5	2,772,266	35,500	1064	4033	310M
id	711,624	3	2,234,953	4,230	395	1317	290M
ar	1,625,651	5	149,043	17,900	558	1681	274M
bn	160,408	3	445	3,270	298	788	273M
pt	1,138,923	4	1,937.959	9,660	596	1935	264M
ru	2,012,648	4	509,503	13,300	799	2307	255M
ur	215,081	3	454	3,220	204	658	232M
de	2,964,125	5	1,227,473	42,900	789	348	133M
ja	1,438,806	1	317,394	38,200	596	2887	123M
mr	98,559	2	275	1,480	193	642	99M
te	101,681	1	13	2,120	223	589	96M
tr	624,742	4	370,727	8,490	398	1389	90M
ta	169,766	3	728	3,980	263	1030	87M
vi	1,294,281	4	44,477	3,010	474	1188	86M
tl	47,891	3	933	1,100	116	451	83M
ko	691,121	4	793,921	16,900	534	2741	82M
ha	51,659	2		823	98	441	79M
jv	74,159	1		535	76	342	68M
it	1,893,522	4	184,535	14,400	516	2129	68M
gu	30,474	1	23	263	174	581	62M
th	169,192	3	41,628	12,700	326	900	61M
kn	33,026	1	143	1540	178	534	59M
am	15,374	2	96	1110	117	493	58M
yo	34,080	2	18	799	123	458	46M

Table 9: Some resources for transfer learning from high resource languages to growth opportunities

the term, *growth*, over terms such as low resources to refer to languages with more speakers than resources, such as many of the languages in Table 9. Table 9 is sorted by the number of speakers.[12] The columns are based on:

- Articles in Wikipedia[13]
- Joshi classification[14] (Joshi et al., 2020)
- Abstracts in Semantic Scholar (S2)
- Articles in ACL Anthology[15]
- Datasets and Models in HuggingFace (HF)

The good news is that we have more resources these days for growth languages than we had for English when we started EMNLP in 1990s. In addition to the resources in Table 9, there is support for most of these languages in multilingual LLMs, Google Translate, and No Language Left Behind (NLLB) (NLLB Team et al., 2022).

How can we transfer between languages with more resources and languages with fewer resources? The crux of the problem is to construct a comparable corpus of English and the growth language. Given that, there are a number of well-established methods to train language models.

Many efforts start by pivoting from English. That is, they use English documents as the source text, and then translate from English to the growth opportunity. Filter bubbles are a problem for this approach. This approach will not learn aspects of the low resource language that go beyond what is in the high resource language.

We suggest using translation in the reverse direction, as well as similarities based on recommender technologies in academic search engines. That is, we will start with source texts in the growth language such as Wikipedia articles and academic papers in Semantic Scholar (S2) (Wade, 2022). We can then find "nearby" English by several means:

1. Translation from growth language to English
2. Similar in a BERT-like vector space using Specter vectors (Cohan et al., 2020) from S2
3. Similar in terms of random walks on citations

By starting with documents in the growth language, we avoid the filter bubble criticism above. In addition, professional translators specialize in one direction and not the other. They prefer to translate into their stronger language than vice versa. We

[12] `https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers`

[13] `https://en.wikipedia.org/wiki/List_of_Wikipedias`

[14] `https://microsoft.github.io/linguisticdiversity/assets/lang2tax.txt`

[15] Based on searches such as `https://aclanthology.org/search/?q=hausa`

suggest similar logic applies to transfer learning. It is better for systems that are stronger in English to translate into English than vice versa.

3.3 Filter Bubbles: A Monolingual Use Case

3.3.1 Filter Bubbles in News and Academia

There are opportunities for CC to burst filter bubbles, both in monolingual and multilingual applications. With the rise of social media and cable news, we all live in filter bubbles. You may remember EMNLP was in Hong Kong just before COVID. I was interested in the coverage of demonstrations in Hong Kong. The story was very simple in New York and in Beijing. The two perspectives disagreed in many respects, of course, but they agreed on simplicity. When I went to Hong Kong for EMNLP, I learned that the story was anything but simple. In short, we all have a tendency to oversimplify the truth, especially about events that are far way, *of which we know little*,[16] like the famous cover of the New Yorker magazine with a view of the world from 9th avenue.[17]

Ground News has created a business by helping people see their blind spots.[18] They track coverage in a range of different news outlets, and report who is saying what. Is this story covered more by outlets on the left or by outlets on the right?

This is an excellent place to start, but the news is fragmented in many more dimensions than just left/right in America. The conflict in Sryia, for example, overlays three dimensions: (1) America/Russia, (2) Sunni/Shia and (3) Turkey/Kurds. More dimensions are more challenging.

Academic conflicts have even more dimensions. Each school of thought has its position, and its friends and foes. In (Church, 2011), I suggested the pendulum has been swinging back and forth between empiricism and rationalism every 20 years. Here is a slightly updated version of that argument:

- Empiricism I (1950s):
  Shannon, Skinner, Harris, Firth
- Rationalism I (1970s):
  Chomsky, Minsky
- Empiricism II (1990s):
  IBM, AT&T Bell Labs, EMNLP, WWW
- Empiricism III (2010s):
  Deep networks, LLMs, chat bots, RAG

Why is the gap around 20 years? One suggestion involves the cliche that grandparents and grandchildren have a natural alliance. Each academic generation rebels against the their teachers. Chomsky and Minsky rebelled against methods that were popular in the 1950s, and my generation returned the favor by reviving those methods. When we started EMNLP (Empirical Methods in Natural Language Processing), the E-word was an act of rebellion.

3.3.2 How can CC burst these filter bubbles?

Suppose we consider Semantic Scholar to be a CC full of multiple overlays that go well beyond Empiricism and Rationalism. We can model the literature as schools of thought with agreements within clusters and disagreements across clusters.

As suggested above, these days, it has become standard practice to represent everything with vectors. We can use vectors to represent papers, as well as schools of thought. Cosines can be used to estimate agreement and disagreement. There are a number of ways to represent papers as vectors. Two suggestions were mentioned above: BERT-like Specter vectors and random walks on citations[19] (Zhang et al., 2019). We normally use comparable corpora in bilingual applications, but this application, clustering, has applications in both bilingual and monolingual settings.

3.3.3 Comparable Corpora and Bots

Web, news and social media offer many different perspectives and points of view. American bots are trained on American corpora; these bots currently lack a historian's ability to approach conflicts from multiple perspectives.

As homework for my NLP class (Church, 2024), I asked students to write essays about the Opium War from multiple perspectives including both the East and the West. They were encouraged to use bots, but were told they would be responsible for the content. I had hoped students would rewrite output from the bots, but few did. Even students from China handed in essays from an American perspective, because American bots are trained on American corpora. These bots do not mention "the century of humiliation,"[20] a perspective in the East which is motivating efforts to compete with the

[16] https://www.iwp.edu/articles/2023/04/18/a-quarrel-in-a-faraway-country-between-people-of-whom-we-know-nothing/

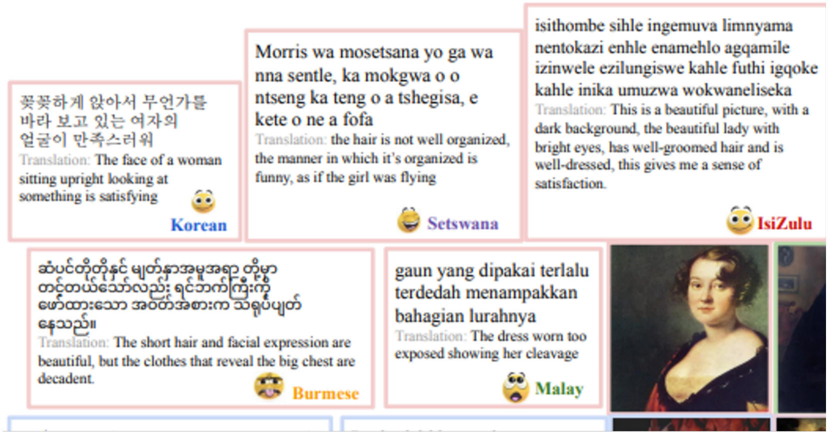
[17] https://en.wikipedia.org/wiki/View_of_the_World_from_9th_Avenue

[18] https://ground.news/blindspotter/methodology

[19] https://github.com/VHRanger/nodevectors

[20] https://www.uscc.gov/sites/default/files/3.10.11Kaufman.pdf

(a) This tree is labeled "sad" in English and Chinese, but annotations in Arabic are more positive.



(b) Many annotations are positive, but some object to the dress as too revealing.

Figure 1: Emotion labels and captions depend on annotator's background (language/culture).

West in AI so China does not fall behind in technology like it did during the Opium Wars.

Bot technology remains far behind historians like Platt (2019). Bots see the world from a single (American) perspective. Filter bubbles are dangerous; they contribute to trade wars and worse.

3.4 Comparable Corpora and Pictures

We normally think of corpora as text, but now that we are representing everything as vectors, we can generalize corpora to include more modalities: text, speech, pictures, speech, video, etc. As mentioned above, we are worried about pivoting from English prompts. If we start with English prompts, then we are likely to bias responses toward an English perspective. Mohamed et al. (2022, 2024) starts with pictures from WikiArt[21] as prompts. Annotators are asked to add emotion labels and captions in 28 languages, as shown in Figure 1. Different annotators label pictures with different emotion labels and captions, depending on their language and background. The papers refer to a GitHub with a benchmark, as well as baseline implementations of captioning systems that transfer from high resource languages to growth opportunities. Hopefully, the community will accept the challenge and come up with even better systems that embrace diversity over many regions, cultures and languages.

It should be possible to beat a baseline system that translates the captions from English to growth languages. Consider the objections to the dress in Figure 1b. This is a case where it should be possible to outperform a captioning system that translates from English because these objections are unlikely to be found in English captions. In fact, a reviewer asked for an ethics review, objecting to the objections to the dress. We are not siding with one annotator over another, but we object to the objection to the objection. It is not appropriate for us to impose American sensibilities on the rest of the world. Rather than remove biases from corpora (and WikiArt), we hope to build bots that will be more aware of regional sensitivities to topics such as: dress, nudes, religion and alcohol.

4 Conclusions

This paper started with a review of the history of comparable corpora in section 2, followed by a discussion of challenges for the future in section 3.

1. BLI is based on a (too) simple view of the lexicon. Can we capture etymology? Differences between monolingual and bilingual senses?
2. Transfer learning to growth languages: Avoid pivoting via English. Better to prompt with pictures. If we have to translate, it is better to translate into English than out of English to avoid imposing American values on others.
3. Similarities between CC and recommender systems for academic search: can we compare and contrast a query document with candidate recommendations? Can we cluster documents in monolingual and multilingual settings, and compare/contrast within and across clusters?
4. Filter bubbles: chat bots currently lack a historian's ability to approach conflicts from multiple perspectives; bots made in America are trained on corpora from an American perspective with American biases. Can we capture "possible worlds" and diverse perspectives?

[21] https://www.wikiart.org/

References

Idris Abdulmumin, Sthembiso Mkhwanazi, Mahlatse Mbooi, Shamsuddeen Hassan Muhammad, Ibrahim Said Ahmad, Neo Putini, Miehleketo Mathebula, Matimba Shingange, Tajuddeen Gwadabe, and Vukosi Marivate. 2024. Correcting FLORES evaluation dataset for four African languages. In *Proceedings of the Ninth Conference on Machine Translation*, pages 570–578, Miami, Florida, USA. Association for Computational Linguistics.

Vesa Akerman, David Baines, Damien Daspit, Ulf Hermjakob, Taeho Jang, Colin Leong, Michael Martin, Joel Mathew, Jonathan Robie, and Marcus Schwarting. 2023. The ebible corpus: Data and model benchmarks for bible translation for low-resource languages. *arXiv preprint arXiv:2304.09919*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

Yehoshua Bar-Hillel. 1960. The present status of automatic translation of languages. *Advances in computers*, 1:91–163.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Sebastian Bruch. 2024. *Foundations of Vector Retrieval*. Springer.

Kenneth Church. 2011. A pendulum swung too far. *Linguistic Issues in Language Technology*, 6.

Kenneth Church. 2024. Emerging trends: When can users trust GPT, and when should they intervene? *Natural Language Engineering*, pages 1–11.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Andreas Eisele and Yu Chen. 2010. MultiUN: A multilingual corpus from united nation documents. In *LREC*.

JR Firth. 1957. A synopsis of linguistic theory 1930-1955. *Studies in Linguistic Analysis*.

Pascale Fung. 2000. A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora. *Parallel Text Processing: Alignment and use of translation corpora*, pages 219–236.

Pascale Fung and Kenneth Ward Church. 1994. K-vec: A new approach for aligning parallel texts. In *COLING 1994 Volume 2: The 15th International Conference on Computational Linguistics*.

Pascale Fung and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 414–420, Montreal, Quebec, Canada. Association for Computational Linguistics.

William A Gale, Kenneth W Church, and David Yarowsky. 1992. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26:415–439.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Zellig Harris. 1964. Distributional structure. *Word*, (2):146–162.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Adam Kilgarriff, Frieda Charalabopoulou, Maria Gavrilidou, Janne Bondi Johannessen, Saussan Khalil, Sofie Johansson Kokkinakis, Robert Lew, Serge Sharoff, Ravikiran Vadlapudi, and Elena Volodina. 2014. Corpus-based vocabulary lists for language learners for nine languages. *Language Resources and Evaluation*, 48(1):121–163.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. *Advances in neural information processing systems*, 27.

Vukosi Marivate, Daniel Njini, Andani Madodonga, Richard Lastrucci, and Jenalea Dzingirai, Isheanesu Rajab. 2023. The vuk'uzenzele south african multilingual corpus.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Youssef Mohamed, Mohamed Abdelfattah, Shyma Alhuwaider, Feifan Li, Xiangliang Zhang, Kenneth Church, and Mohamed Elhoseiny. 2022. ArtELingo: A million emotion annotations of WikiArt with emphasis on diversity over language and culture. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8770–8785, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Youssef Mohamed, Runjia Li, Ibrahim Said Ahmad, et al. 2024. No culture left behind: ArtELingo-28, a benchmark of WikiArt with captions in 28 languages. In *EMNLP*, pages 20939–20962, Miami, Florida, USA. Association for Computational Linguistics.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

NLLB Team, Marta Ruiz Costa-jussà, James Cross, Onur Celebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Alison Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon L. Spruit, C. Tran, Pierre Yves Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzm'an, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *ArXiv*, abs/2207.04672.

Stephen R Platt. 2019. *Imperial Twilight: The Opium War and the End of China's Last Golden Age*. Vintage.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 320–322, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Philip Resnik and Noah A. Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Peter H Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.

Serge Sharoff, Reinhard Rapp, and Pierre Zweigenbaum. 2023. *Building and Using Comparable Corpora for Multilingual Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Springer Nature.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Alex D Wade. 2022. The semantic scholar academic graph (s2ag). *Companion Proceedings of the Web Conference 2022*.

Jie Zhang, Yuxiao Dong, Yan Wang, Jie Tang, and Ming Ding. 2019. Prone: Fast and scalable network representation learning. In *IJCAI*, volume 19, pages 4278–4284.