Do we Name the Languages we Study? The #BenderRule in LREC and ACL articles

Fanny Ducel, Karën Fort, Gaël Lejeune, Yves Lepage

Do we Name the Languages we Study? The #BenderRule in LREC and ACL articles

Fanny Ducel[†], **Karën Fort**[†⋆], **Gaël Lejeune**[†], **Yves Lepage**[‡]
[†]STIH, Sorbonne Université, France ⋆ Université de Lorraine, CNRS, Inria, LORIA, France, [‡]Waseda University, Japan
ducelfanny@gmail.com, karen.fort@loria.fr, gael.lejeune@sorbonne-universite.fr, yves.lepage@waseda.jp

Abstract

This article studies the application of the #BenderRule in Natural Language Processing (NLP) articles according to two dimensions. Firstly, in a contrastive manner, by considering two major international conferences, LREC and ACL, and secondly, in a diachronic manner, by inspecting nearly 14,000 articles over a period of time ranging from 2000 to 2020 for LREC and from 1979 to 2020 for ACL. For this purpose, we created a corpus from LREC and ACL articles from the above-mentioned periods, from which we manually annotated nearly 1,000. We then developed two classifiers to automatically annotate the rest of the corpus. We show that LREC articles tend to respect the #BenderRule (80 to 90% of them respect it), whereas only around half of ACL articles do. Interestingly, over the considered periods, the results appear to be stable for the two conferences, even though a rebound in ACL 2020 could be a sign of the influence of the blog post about the #BenderRule.

Keywords: #BenderRule, language diversity, ethics

1. Introduction

1.1. Background

This article is positioned in the field of NLP4NLP, an acronym known from the series of articles by Mariani et al. (2019a) (part II in (Mariani et al., 2019b)). NLP4NLP consists in applying Natural Language Processing (NLP) techniques and exploring the field of NLP itself, by mainly inspecting the content of published articles. The aims of NLP4NLP can be various. For example, exploring the progress in the field, e.g., identifying discoveries or hot topics (Mariani et al., 2013; Mariani et al., 2014; Buitelaar et al., 2014), addressing ethical issues, e.g., reuse and plagiarism (Mariani et al., 2016)[1], or even questioning the reliability of the methods used, e.g., in machine translation (Marie et al., 2021)[2]. This article addresses a question with both ethical and scientific implications: we inspect the number and which languages are studied in two NLP conferences, LREC and ACL, that have already been contrasted in a previous work (Buitelaar et al., 2014). To that end, we firstly annotate a sample of articles by hand and then generalize our results to the entire collection of articles by using linguistic feature extraction and classification techniques from machine learning.

1.2. Topic Addressed in this Article

The problem of *the number and which languages* are studied in the field of NLP has been raised recently, although it had been identified earlier in the history of grammar or linguistics (see Section 1.3). As for NLP, even if the majority of research is still devoted to a limited number of languages, the situation seems to be evolving. This is at least what Joshi et al. (2020) suggest in their article on linguistic inclusion and diversity in the scientific community. Being concerned with language diversity implies that the languages studied *matter* and therefore should at the very least be *stated*. This might sound obvious, but, a decade ago, Bender (2011) noticed that many researchers simply do not cite the languages they work on. She consequently urged researchers to specify the name of the languages being studied, even (or especially) when it is English:

> "Do state the name of the language that is being studied, even if it's English. Acknowledging that we are working on a particular language foregrounds the possibility that the techniques may in fact be language specific. Conversely, neglecting to state that the particular data used were in, say, English, gives [a] false veneer of language-independence to the work." (Bender, 2011, p. 18)

Eight years later, in a blog post (Bender, 2019), she reminded the community that "English is neither synonymous with nor representative of natural language". Specifying the languages studied thus became the #BenderRule, which can be summarized as follows:

> "Always name the language(s) you're working on." (Bender, 2019)

Indeed, the non application of the #BenderRule raises linguistic, sociological and ethical issues. Not naming the language studied (which is often the case for English), implicitly implies that the methods developed could work on any other language, as if English were a neutral and universal language. However, as argued

[1]With the felicitous conclusion, at that time, that "plagiarism is very uncommon in our community."

[2]With guidelines recommending the use of more than one evaluation metric, and the systematic computation of statistical significance, to cite only the first two recommendations.

by Bender herself, English is far from representing the diversity of languages.

1.3. Linguistic Issues

Examples of now well-identified biases towards something believed universal at some point in time can be given, from the history of grammar and linguistics. Let us take the example of word order. In the European history of grammar, the very influential *Grammaire de Port-Royal* (Arnauld and Lancelot, 1662) took the order of words in French as the "natural" one. It claimed its universality, with the consequence that the order in other languages, like German, had to be explained by casting it back to that of French. Fixity of word order is one of several traits listed by Bender to insist on the fact that English is not universal ("English has relatively fixed word order"). Earlier, in linguistics, it had already been pointed at as somewhat special. Specifically, Mel'čuk (1988, p. 4) warned that focusing on it might well lead to methodological issues:

> English is very exotic in that it uses constituency as its only expressive device in syntax, i.e., as the only device for encoding syntactic structure in actual sentences. [...] Constituency is a MANIFESTATION of syntactic structure, not syntactic structure itself. But thinking or even simply working mostly with English lures the researcher into mistaking this idiosyncratic surface trait of a particular language [...] for a universal mechanism of syntactic representation.

With the unfortunate habit of citing articles no older than a decade back, there exists a risk of forgetting about the (even relatively recent) history of sciences, and consequently a risk of repeating methodological mistakes from several decades or even centuries ago.

1.4. Sociological and Ethical Issues

In an article on meta-learning, Nooralahzadeh et al. (2020) take note of the fact that

> "although there is growing awareness in the field, as evidenced by the release of datasets such as XNLI (Conneau et al., 2018), most NLP research still only considers English (Bender, 2019)."

Reducing studies to only English might have sociological impacts and real consequences on the diversity of NLP works. This leads to a vicious circle: working only on English becomes more rewarding (i.e., chances to have an article accepted are higher) because "[w]ork[ing] on languages other than English is often considered 'language specific' and thus reviewed as less important than equivalent work on English." (Bender, 2019). As stated by Hovy and Spruit (2016),

> "[p]otential consequences are exclusion or demographic misrepresentation. This in itself already represents an ethical problem for research purposes, threatening the universality and objectivity of scientific knowledge (Merton, 1973)."

Along the same line, Bender et al. (2021) claim that "most language technology is built to serve the needs of those who already have the most privilege in society". Indeed, scientific research can result in the development of products, applications and software that will only be accessible to English (and, to a lesser extent, Chinese) speakers. Such concerns of unequal treatment of languages have been raised in various places. For instance, Wagner (2021) uses Bender's words in the introduction of her thesis on Icelandic to argue that NLP plays a role in the extinction or survival of languages.

1.5. Contributions of this Article

The extent to which the #BenderRule is applied or not is unclear and, to our knowledge, there is no published work that explicitly addresses this issue[3], apart from some studies focusing on linguistic diversity and representativeness, like (Joshi et al., 2020) already mentioned above. This article is a first attempt at quantifying the application of the #BenderRule, i.e., to empirically examine whether and to what extent the #BenderRule is actually applied. To this end, we train two classifiers to automatically annotate research articles with whether they apply the #BenderRule, and which language(s) they study. The study reported in this article is conducted on two corpora in the field of NLP: articles from the successive occurrences of Language Resources and Evaluation Conference (LREC) and the Annual Meetings of the Association for Computational Linguistics (ACL).[4]

The contributions of this article are as follows:

- we built a corpus of nearly 14,000 articles from the LREC and ACL conferences;
- we manually annotated a sample of nearly 1,000 of these articles, i.e., we checked whether the #BenderRule is applied or not, and took record of the studied languages when possible;
- we trained classifiers on these manual annotations and applied them on the entirety of the articles in both conferences;
- we analyzed the results from a contrastive and diachronic perspective and extracted general trends.

We detail the methodology used to build the corpus and the classifiers in the following section (Section 2). We analyze the results obtained in Section 3. Finally, we discuss the limits of our work in Section 4.

[3] For instance, the impressively extensive study of the LREC archive in (Mariani et al., 2014) mentions small communities working on specific languages, but does not provide any figure about the languages studied in LREC, nor does it give the names of these specific languages.

[4] The #BenderRule tells us that it does not go without saying that the language of the corpora is English.

2. Methodology

2.1. Corpus Building

We created a corpus of articles in English from LREC (from 2000 to 2020) and ACL (from 1979 to 2020). To do so, we converted all the PDF articles from the LREC conferences (6,715 files) to text files using Dallas Card's scripts[5]. We obtained the PDF files from the official website of the LREC conference[6] and from the ACL Anthology[7] for the articles published after 2016. Some of the articles from ACL were already in text format. We obtained the files from 1979 to 2016 from the AAN Anthology Network Corpus (Radev et al., 2013)[8]. We performed a pdftotxt conversion on the ones from after 2016 using the scripts mentioned above. This ACL sub-corpus (text+pdftotxt) contains 7,262 articles.

We then eliminated the articles which presented OCR issues and could not be used (46 LREC files and 24 ACL files). We detected these files thanks to a very simple string-matching test. The text files that do not include the token "the" can be considered unreadable due to OCR issues. Finally, we obtained a corpus of 13,931 articles, with two sub-corpora, one of 6,669 LREC articles and one of 7,262 ACL articles (see Table 1).

	LREC	ACL
Text files	0	4,867
Converted files	6,715	2,419
Excluded files	46	24
Total	6,669	7,262

Table 1: Composition of the corpus.

2.2. Manual Annotation

One of the authors of the present article manually annotated 550 randomly selected LREC articles (50 per edition of the conference, representing approximately 8% of the LREC sub-corpus) and 420 randomly selected ACL articles (10 per year, representing approximately 5% of the ACL sub-corpus), for a total of 970 articles. Each article was annotated with one of the four following categories, along with the languages studied:

- Applied: The #BenderRule is applicable and applied. The language studied is clearly mentioned somewhere in the article.
- Deducible: The #BenderRule is applicable, not applied but deducible. That is, there is at least one monolingual resource that is mentioned, which is listed in the LREMap, thus helping the annotator in deducing the language(s) studied.[9]
- Non-deducible: the #BenderRule is applicable, not applied *and* not deducible, i.e., the language is not mentioned and cannot be deduced from a resource name.[10]
- Non-applicable: The #BenderRule is not applicable. For instance, many theoretical articles do not focus on any particular language and even if they may use English examples, the claims are usually of a more general nature.

We proceeded in the following way for the establishment of the four above categories. We first performed trial annotations, then identified the problems and consequently designed annotation guidelines, so as to ensure a consistent annotation. The three following rules were established:

- i) an article is annotated as "Applied" when it mentions the studied language, even if this mention appears only once at the end of the article,
- ii) the LREMap (Calzolari et al., 2012)[11] was selected as the reference for the "Deducible" class[12],
- iii) we decided to exclude using the examples given in the article to deduce the language, as they can be in languages neither we nor a classifier can identify and because they are not reliable enough as they can be translations (usually into English) from the studied language.

The distribution of the different categories in the manual annotation for LREC and the ACL sub-corpora is illustrated in Figure 1. It shows that LREC articles apply the #BenderRule much more often than ACL articles: 90% of LREC articles in our sub-corpus apply the #BenderRule, against 55% in ACL. Said in another way, and following the definitions given above for the four categories, 45% of ACL articles in our manually annotated corpus do not cite any language. Now, considering that 13% in these 45% of articles have been categorized as "Non-applicable", there are 32% articles in which the #BenderRule has not been applied when it should have. This represents almost a third of the annotated ACL articles and cannot be considered negligible.

[5] www.github.com/dallascard/acl-papers
[6] See: https://www.lrec-conf.org/.
[7] See: https://aclanthology.org/.
[8] See: https://aan.how/.

[9] The use of a "deducible" category is disputable. We acknowledge that letting the language be inferred is different from mentioning it explicitly. Therefore, the #BenderRule can be considered unenforced in such cases, and the impact in terms of visibility of the studied language is not as important as when the #BenderRule is properly enforced.

[10] As counter-examples, it is obvious that the Wall Street Journal (WSJ) is for (American) English.

[11] See: http://lremap.elra.info/.

[12] We used this reference because the idea was to put ourselves in a reader's shoes who doesn't have access to databases such as ELRA's

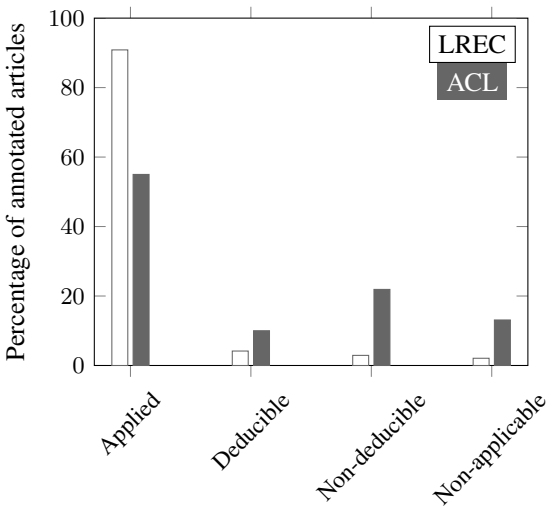

Figure 1: Results of the manual annotation of the articles of LREC and ACL, in percentage (LREC in white, ACL in black).

This has to be compared with (100% − 90%) − 2% = 8% for LREC articles. To summarize this comparison, four times more articles do not apply the #BenderRule when they should in ACL than in LREC.

To check the quality and consistency of the annotations, the three other authors double-annotated a selection of the above articles, as shown in Table 2, so that approximately 7% of the annotations were duplicated. The sampling was generally random, but we tried to balance the categories (the entire corpus itself is obviously not balanced) and we added some complicated cases on purpose. For LREC, three to four articles per conference were double-annotated and for ACL, one article per conference between 2000 and 2020, except for 2000, 2005, 2010, 2015 and 2020 for which two articles were in the selection. Moreover, one article every two conferences between 1980 and 1998 was double-annotated too. Duplicated annotations allowed the computation of the inter-annotator agreement. We computed the observed inter-annotator agreement between annotator 0 (A_0, who manually annotated almost 1,000 articles) and each of the others. We obtain an agreement of 80% with annotator 1, 83% with annotator 2 and 63% with annotator 3. This last result can probably be explained by the fact that annotator 3 annotated less articles than the others, so the articles they got were especially complicated and the classification was questionable. If we merge the annotations from annotators 1, 2 and 3 and compare them to annotator 0, we obtain a Cohen Kappa (Cohen, 1960) of 0.82, which can be considered good (Artstein and Poesio, 2008).

To extend the study and cover all available articles in both conferences, we developed two classifiers.

2.3. Classification using Pattern-Matching

A first, baseline classifier, Class_{PM} (PM for pattern-mactching), relies on an approach based on string-matching.

	LREC	ACL	Total	Agreement with A_0
Annotator 1	20	10	30	80.0%
Annotator 2	11	16	27	82.5%
Annotator 3	6	10	16	62.5%
Total	37	36	73	

Table 2: Double-annotated articles per annotator and conference, and observed agreement with A_0.

We consider that an article applies the #BenderRule when at least one language name is found in the text. For that, we use a list of over 500 different ISO 639-2 language names in English[13]. For each individual article, we memorize the languages mentioned, along with their number of mentions. The languages are obtained by pattern-matching after standard preprocessing of the texts, i.e., removal of punctuation, lowercasing, and tokenization. If at least one of the tokens matches one of the language names in the list, the article is classified as applying the #BenderRule ("Applied").

The above method allows us to put aside the articles in which no language is cited. This might be because the authors did not apply the #BenderRule, or because the #BenderRule cannot be applied (theoretical or meta-linguistic work). We consider an article as "Deducible" if it contains at least one name of a language resource listed in the `LREMap`.[14] The list of all the monolingual language resources present on the `LREMap` website was extracted by using Web scraping methods based on the Python library `BeautifulSoup` with various improvements[15]. Tokens in articles are matched against language resource names and paths. Articles citing at least one resource from the list are put in the "Deducible" category.

The remaining articles are classified as not applying the #BenderRule or the #BenderRule is not applicable (referred to as "N/A" in Table 3 and Figure 2)[16]. Note that the classifiers cannot differentiate between the articles based on linguistic data which do not cite the studied language (i.e. that should apply the #BenderRule but do not) from those that do not have to mention any language, for example because the topic is theoretical or meta-linguistic.

[13] See: `https://github.com/ISO639/2`.

[14] We added to the list the following language resources, which we consider popular enough (for English and French): "Le Monde, France info, France inter, NYT, PropBank, Washington Post, Wall Street Journal, Brown Corpus, Le Robert, Switchboard".

[15] See T. Ujhelyi's tutorial: `https://data36.com/scrape-multiple-web-pages-beautiful-soup-tutorial/`.

[16] It should be noted that the number of classes in the automatic processing therefore becomes three, to be compared with four during manual annotation.

2.4. Classification using Machine Learning

The approach based on string-matching does not recognize the cases where a language is mentioned without being the one actually studied. For instance, authors tend to mention the languages they would like to study in future works in the conclusion, or the languages that have already been studied by colleagues, in the introduction or the state-of-the-art sections.

Thus, we train a sentence classifier with the aim of correcting the above-mentioned issues. This sentence classifier uses supervised machine learning techniques. It is trained on 2,625 sentences extracted from the manually annotated ACL articles and which contain at least one language name. Those sentences coming from articles that were manually annotated as "Applied" were categorized as such, whereas the ones coming from articles that were considered as not applying the #BenderRule were annotated as not applied or not applicable, i.e., "N/A". We use CountVectorizer from `scikit-klearn` to vectorize the corpus, with default parameters, and LogisticRegression (with the saga solver) as the supervised classification algorithm as this proved to be the best combination. Using tf·idf weighting or working with n-grams longer than unigrams did not improve the results. Logistic Regression showed the best results among the classifiers that we tested (Decision Trees, Random Forests, Bayesian and SVM classifiers).

The classifier for articles, Class_{ML} (ML for machine learning), classifies an article as "Applied" if it contains at least one sentence classified as such by the sentence classifier. Typically, a sentence like "The morphological status of affixes in Chinese [...]" (Tseng et al., 2020) would be classified as "Applied", whereas "These annotations can demarcate components of signs, [...] or a translation into another language like English." (Bragg et al., 2019) would not be classified as "Applied".

3. Results

3.1. Performance of the Classifiers

We evaluate our classifiers by comparing their outputs with the manual annotation we performed. The results are presented in Table 3.

Class_{ML} performs better than Class_{PM} on ACL articles in terms of precision and presents a better Spearman correlation with Annotator 0. However, their performance is comparable on LREC articles. Results are less accurate for the category "Deducible" for ACL articles. We could question our definition of "Deducible" and especially our choice to use the `LREMap` as the sole reference, as a number of corpora do not appear in it. Nonetheless, it is difficult to assert whether a resource can be considered reliable enough or not and the `LREMap` has the merit of being a well-established meta-reference.

Furthermore, some articles are missing from the category "Deducible" because they remain falsely classified as applying the #BenderRule ("Applied").

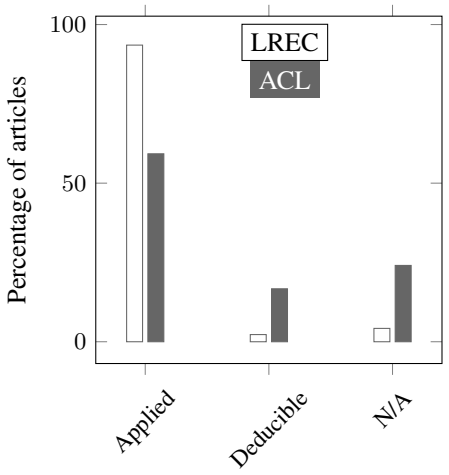

Figure 2: Results, in percentage, of the automatic classification by Class_{ML} of LREC and ACL articles into three categories: #BenderRule applied (Applied), language(s) studied deducible (Deducible), or none of the previous (N/A). LREC in white, ACL in black.

3.2. Results of the Experiments

3.2.1. Classification Results

Class_{ML} classifies 40% of all the ACL articles in our corpus as not applying the #BenderRule, which means that they either do not mention any language, or mention some, which are not the ones studied. Similarly, among the 420 manually annotated articles, 45% do not apply the #BenderRule according to Class_{ML}. We explain this high proportion by the longer period of time for ACL (1979–2020, 2000–2020 for LREC). Multilingual works were rarer in the past, it was maybe obvious to work on English at that time, and therefore authors did not specify the studied language.

Articles from LREC seem to apply the #BenderRule more often. Class_{PM} predicts that only 17% of articles do not apply it (14% according to Class_{ML}), as compared to 9% in manual annotation.

The difference between the two conferences can be explained by the fact that LREC, following the aims of the conference, presents more linguistic diversity. Languages that are usually less represented are (almost) always explicitly mentioned. We note that in the last category (N/A), there are more articles where the #BenderRule is not applicable than not applied. Therefore, even when there is no language mentioned, it is legitimate. Furthermore, we noticed during manual annotation that, in some LREC articles, not only the language was specified but also its particular variant.

Figure 2 shows the distributions of articles obtained by automatic classification using Class_{ML} over the whole LREC and ACL sub-corpora. They reasonably reproduce the distributions obtained by manual annotation of a sample of articles in each sub-corpus (see Figure 1). For ACL, the manual annotation gives 55% of "Ap-

	Class_{PM}							
	LREC (correlation = 0.634)				ACL (correlation = 0.509)			
	Prec.	Recall	F-meas.	# instances	Prec.	Recall	F-meas.	# instances
Applied	0.894	**0.993**	0.941	440	0.729	**0.978**	0.835	231
Deducible	**1.000**	**0.650**	**0.788**	20	**1.000**	0.595	**0.746**	42
N/A	**0.792**	0.422	0.551	90	0.741	0.429	0.543	147
Macro avg.	**0.895**	0.688	0.760	550	0.823	**0.667**	**0.708**	420
	Class_{ML}							
	LREC (correlation = **0.671**)				ACL (correlation = **0.741**)			
	Prec.	Recall	F-meas.	# instances	Prec.	Recall	F-meas.	# instances
Applied	**0.908**	0.986	**0.946**	440	**0.856**	0.974	**0.911**	231
Deducible	**1.000**	0.600	0.750	20	**1.000**	0.286	0.444	42
N/A	0.767	**0.511**	**0.613**	90	**0.752**	**0.741**	**0.747**	147
Macro avg.	0.892	**0.699**	**0.770**	550	**0.869**	0.667	0.701	420

Table 3: Performance of the classifiers and Spearman correlation with Annotator 0 on the two sub-corpora (N/A: not applied or not applicable), the best result for each class and each measure is in bold.

.

plied" vs. 59% predicted, 10% of "Deducible" vs. 17% predicted, and 35% remaining vs. 24% with automatic prediction. "Deducible" is thus doubled in automatic prediction, which mechanically reduces the percentage in the last class.

Manual inspection reveals that some articles are mistakenly assigned to the class "Deducible", because some terms are erroneously considered resources names, some acronyms are polysemous, or because some authors cite the resources used in previous works.

As for LREC, automatic classification yields 94% of "#BenderRule Applied" (vs. 91%), 2% of "Deducible" (vs. 4%) and 4% for "N/A" (vs. 5%).

Among the articles applying the #BenderRule, we notice that some only quote one language once. For ACL, it concerns 13% of the total of articles, against 4% of the manually annotated ones. LREC stands out with only 5% of articles containing a single mention of a language. As for the manually annotated articles, it concerns only 1% of them. We conclude that authors focus more attention on the languages they study in the LREC conference.

3.2.2. Diachronic Study

We show in Figure 3 the proportion of articles applying the #BenderRule per year, for each year of the studied period of time, in each conference.

There does not seem to be any chronological evolution in ACL, the proportion being almost always between 50% and 60%. Nonetheless, we observe an increase in 2020, after a decreasing trend in average from 1986 to 2019. One can ask whether the increase was caused by Bender's blog post in 2019.

For LREC, the proportion of articles applying the #BenderRule exhibits a slowly increasing trend over the years, especially between 2000 and 2010. However it is worth noticing that, the lowest number, which concerns 2000, is already above 80%, that is, greater than the maximum for ACL, reached in 1986 with 68%. This confirms our hypothesis that, LREC being a conference focusing on language resources, the languages being studied are most of the time mentioned, even if it is English (Figure 5 supports this observation).

3.2.3. Languages for which the #BenderRule is not Applied

Figure 4 gives the number of articles per number of languages mentioned. It shows that a large number of ACL articles do not cite any language at all (0 languages): they match the "Deducible" and "N/A" (not applied or not applicable) categories.

Among the articles which apply the #BenderRule, we notice that the languages are more cited when the study is not monolingual. Indeed, it is necessary to name and distinguish the different studied languages in order to compare them in a multilingual work. Figure 4 makes the proportion of cross-lingual works visible, since most multilingual studies apply the #BenderRule. For ACL, the system detects many more articles which do not apply the #BenderRule, the proportion of articles that do not cite any language is almost three times higher than the others. We can also see that monolingual studies are a little more represented than bilingual ones, and that trilingual and quadrilingual studies are even less present. Nonetheless, there are also articles that study up to 13 languages simultaneously.

Figure 4 shows that in LREC, there are significantly less articles which do not cite any language than articles that cite any other number of languages. We even notice that there are more bilingual studies than monolingual ones. We can generally say that LREC articles are much more multilingual, with studies that can concern up to several dozens of languages at once.[17]

Figure 5 shows that the top 10 languages in both conferences are very similar, with only one language be-

[17] During manual annotation we even came across an article dealing with 107 languages, listed by their ISO code, not their names, so that the classifiers could not identify them.

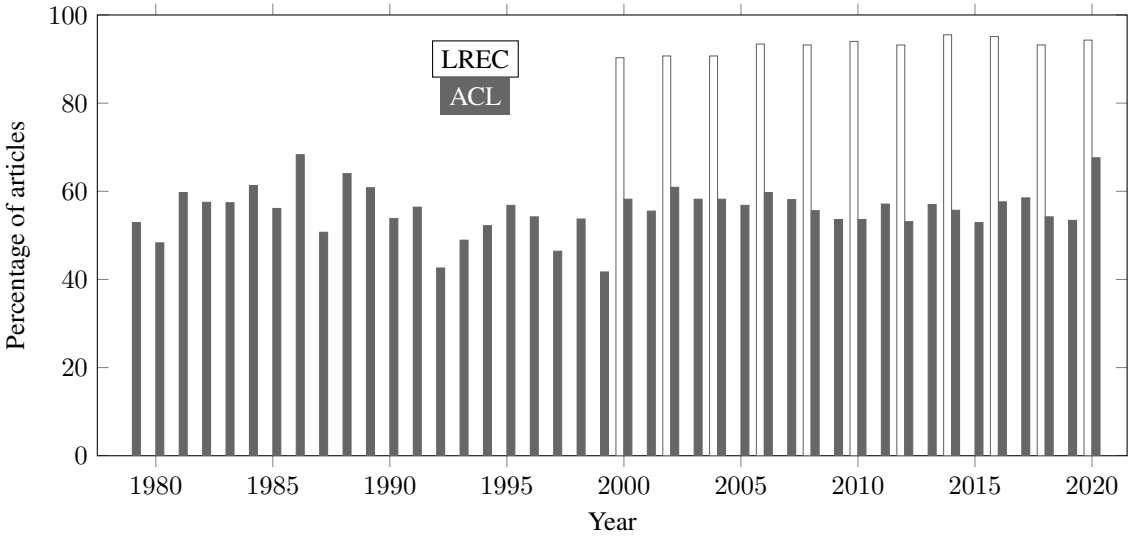

Figure 3: Number of articles applying the #BenderRule, in percentage in each edition of both conferences (LREC in white, ACL in black). LREC is held every two years and its proceedings are only available from 2000.

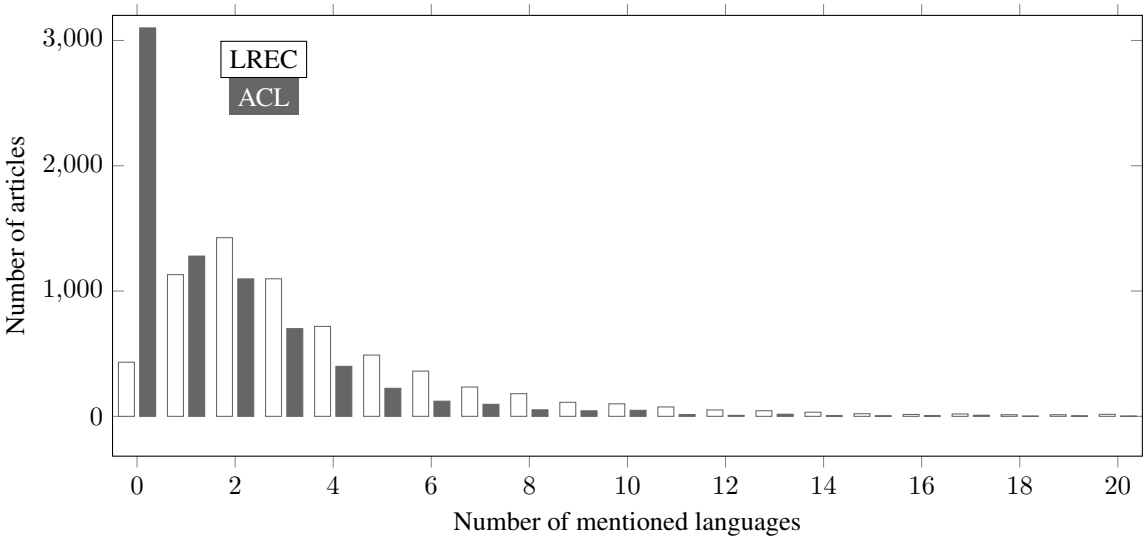

Figure 4: Distribution of number of articles per number of languages mentioned (LREC in white, ACL in black). There is a long tail after 20 languages for both conferences, the number of articles is then almost always one article only. The maximal number of languages is 92 languages for ACL and 90 for LREC.

ing different (Korean is only present in ACL, while Farsi appears only in LREC). Furthermore, it should be noticed that these top languages are spoken in politically or economically important countries or have a relatively large number of speakers. As Bender already noticed, "many of the languages included are close relatives of each other" (Bender, 2009). Indeed, except for Chinese, Japanese and Korean, they all (including Farsi) belong to the Indo-European language family. Moreover, only three languages constitute more than 10% of the total of mentioned languages.

In the LREC corpus, German is the most cited language, while English constitutes less than 16% of the mentions. However, the majority class is in fact the one composed of the 239 other languages present in the corpus. This shows once again the large linguistic diversity offered by LREC.

As for ACL, we observe that Mandarin Chinese comes right after English, which reflects the fact that it is gradually gaining ground in the English-speaking and the scientific community in general. We anticipate that, in a few years, Mandarin Chinese may overtake English in the rankings. This does not mean that Mandarin Chinese will be more studied than English, but that it will be more *mentioned*. The results are of course biased because the non-application of the #BenderRule leads to some invisibility of the word "English" in articles.

If we extract the list of studied languages in the articles which do not apply the #BenderRule (corresponding to languages that are studied without being men-

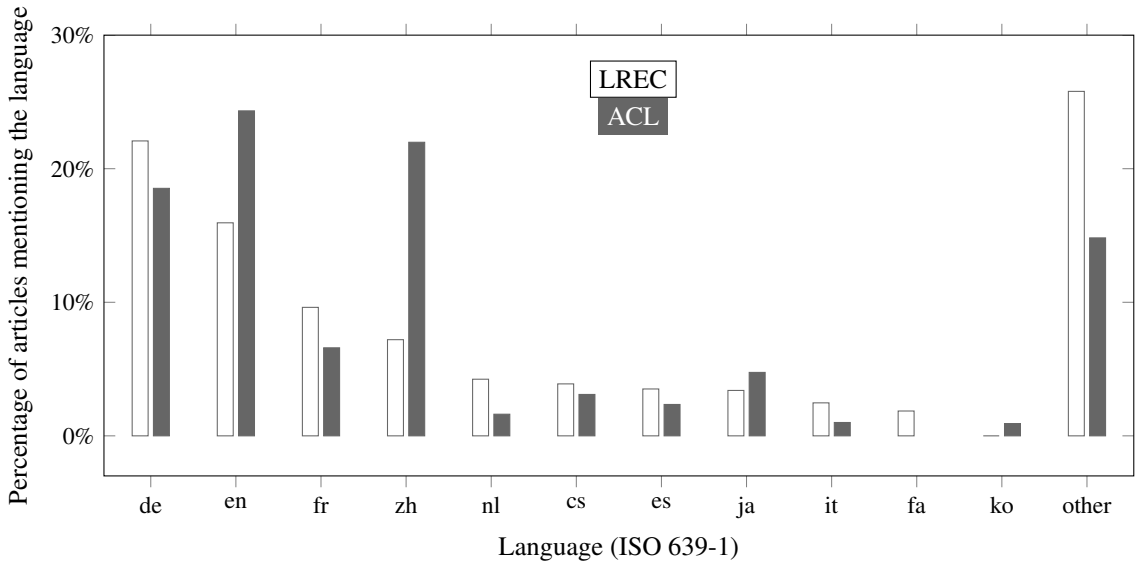

Figure 5: Percentage of articles mentioning a given language in each corpus, over all articles applying the #BenderRule (LREC in white, ACL in black). The languages are sorted by decreasing order of percentage in LREC (white bars). Only the 10 most mentioned languages in each corpus are shown (11 languages in total).

tioned), we only get English. This confirms Bender's assertions that English remains considered as a "default language", "synonymous of natural language" and that there is, indeed, a real "habit of failing to name the language studied when it is English" (Bender, 2019).

4. Potential Limitations of the Study

Obviously, we could not read the entirety of all the articles that we manually annotated, therefore some annotations might be incorrect or questionable. However, the inter-annotator agreement results that we obtained are reassuring.

Moreover, the list of languages that we used is non-exhaustive and we probably missed some. We identified such a case: Yu et al. (2016) worked on 107 languages but only list them with their ISO codes which we do not take into account and could not detect.

Another issue that we did not take into account is the lack of precision concerning the language, even when mentioned. For example, when Chinese is mentioned, the authors should precise which Chinese (usually, as Bender puts it in her blog post, it is Mandarin Chinese), the same goes for almost all languages.

More importantly, we had to exclude 70 articles due to OCR issues (unreadable files). As we did not manually correct the OCR results, there might be some remaining issues that we could not identify.

Finally, we limited our study to articles up to 2020 (the last LREC). The #BenderRule was worded in 2019, that is to say, shortly before that. Therefore, there might be an influence of this proximity that is difficult to assess properly now. In order to evaluate the impact of the #BenderRule on the longer term, these experiments should be rerun a few years from now. However, our main goal was not to evaluate the impact of the blog post, but to assess the prevalence of the issue, hence, this does not influence our research as much.

5. Conclusion

We compared the application of the #BenderRule in two important conferences. We found that LREC articles mention much more the languages they study. We also found that LREC presents more linguistic diversity, more multilingual works, and that, in general, it puts more emphasis on the variety of languages than ACL does. We performed the same experiments using the same programs (with a few adaptations for French) on the TALN conference and obtained results that are more similar to the figures presented here for ACL. We diachronically examined the evolution of the application of the #BenderRule but did not observe any significant change over time. Finally, in both corpora, we observed that when the #BenderRule is not applied, the studied language is English, which confirms some of Bender's assertions.

The code and resources used are freely available on GitHub[18] for replication purposes. In order to assess the impact of Bender's blog post, the classifiers should be re-run in the coming years on the ACL (2021 and following) and LREC (2022 and following) proceedings. Another perspective is to replicate the experiment on the articles from other international and national NLP conferences. As mentioned above, we already started to work on the proceedings of the French national conference TALN and expect to publish the results rapidly.

[18]See: `https://github.com/FannyDucel/bender-rule-lrec-acl`.

6. Bibliographical References

Arnauld, A. and Lancelot, C. (1662). *Grammaire générale et raisonnée de Port-Royal*. chez Prault fils l'aîné, Paris, subsq. ed. 1784.

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Bender, E. M. (2009). Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32, Athens, Greece, March. Association for Computational Linguistics.

Bender, E. M. (2011). On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, 6(3), October.

Bender, E. (2019). The #BenderRule: On naming the languages we study and why it matters. *The Gradient*.

Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., Caselli, N., Huenerfauth, M., Kacorri, H., Verhoef, T., Vogler, C., and Ringel Morris, M. (2019). Sign language recognition, generation, and translation: An interdisciplinary perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, page 16–31, New York, NY, USA. Association for Computing Machinery.

Buitelaar, P., Bordea, G., and Coughlan, B. (2014). Hot topics and schisms in NLP: Community and trend analysis with saffron on ACL and LREC proceedings. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2083–2088, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Hovy, D. and Spruit, S. L. (2016). The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany, August. Association for Computational Linguistics.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (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, July. Association for Computational Linguistics.

Mariani, J. J., Paroubek, P., Francopoulo, G., and Delaborde, M. (2013). Rediscovering 25 years of discoveries in spoken language processing: a preliminary analysis of the ISCA archive. In *Annual Conference of the International Speech Communication Association*, Lyon, France, January.

Mariani, J., Paroubek, P., Francopoulo, G., and Hamon, O. (2014). Rediscovering 15 years of discoveries in language resources and evaluation: The LREC anthology analysis. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Mariani, J., Francopoulo, G., and Paroubek, P. (2016). A study of reuse and plagiarism in speech and natural language processing papers. In *Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL)*, pages 72–83, June.

Mariani, J. J., Francopoulo, G., and Paroubek, P. (2019a). The NLP4NLP Corpus (I): 50 Years of Publication, Collaboration and Citation in Speech and Language Processing. *Frontiers in Research Metrics and Analytics*, 3:1–30.

Mariani, J. J., Francopoulo, G., Paroubek, P., and Vernier, F. (2019b). The NLP4NLP Corpus (II): 50 Years of Research in Speech and Language Processing. *Frontiers in Research Metrics and Analytics*, 3:1–30.

Marie, B., Fujita, A., and Rubino, R. (2021). Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7297–7306, Online, August. Association for Computational Linguistics.

Mel'čuk, I. A. (1988). *Dependency Syntax: Theory and Practice*. State University of New York Press, New York.

Nooralahzadeh, F., Bekoulis, G., Bjerva, J., and Augenstein, I. (2020). Zero-shot cross-lingual transfer with meta learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4547–4562, online, November. Association for Computational Linguistics.

Tseng, Y.-H., Hsieh, S.-K., Chen, P.-Y., and Court, S. (2020). Computational modeling of affixoid behavior in Chinese morphology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2879–2888, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Wagner, A. (2021). *Country, Nation, and Language –*

Machine Translation in Iceland. Ph.D. thesis, Université d'Islande – Faculté des sciences humaines, Juin.

Yu, Z., Mareček, D., Žabokrtský, Z., and Zeman, D. (2016). If you Even don't have a bit of Bible: Learning delexicalized POS taggers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 96–103, Portorož, Slovenia, May. European Language Resources Association (ELRA).

7. Language Resource References

Calzolari, Nicoletta and Del Gratta, Riccardo and Francopoulo, Gil and Mariani, Joseph and Rubino, Francesco and Russo, Irene and Soria, Claudia. (2012). *The LRE Map. Harmonising Community Descriptions of Resources*. European Language Resources Association (ELRA).

Radev, Dragomir R. and Muthukrishnan, Pradeep and Qazvinian, Vahed and Abu-Jbara, Amjad. (2013). *The ACL anthology network corpus*. Springer Netherlands.