Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3158–3163
Marseille, 20-25 June 2022

BaSCo: An Annotated Basque-Spanish Code-Switching Corpus for Natural Language Understanding

Maia Aguirre, Laura García-Sardiña, Manex Serras, Ariane Méndez, Jacobo López
Vicomtech Foundation, Basque Research and Technology Alliance (BRTA)
Parque Científico y Tecnológico de Gipuzkoa, Paseo Mikeletegi 57, Donostia / San Sebastián (Spain)
{magirre, lgarcias, mserras, amendez, jlopez}@vicomtech.org

Abstract

The main objective of this work is the elaboration and public release of *BaSCo*, the first corpus with annotated linguistic resources encompassing Basque-Spanish code-switching. The mixture of Basque and Spanish languages within the same utterance is popularly referred to as *Euskañol*, a widespread phenomenon among bilingual speakers in the Basque Country. Thus, this corpus has been created to meet the demand of annotated linguistic resources in Euskañol in research areas such as multilingual dialogue systems. The presented resource is the result of translating to Euskañol a compilation of texts in Basque and Spanish that were used for training the Natural Language Understanding (NLU) models of several task-oriented bilingual chatbots. Those chatbots were meant to answer specific questions associated with the administration, fiscal, and transport domains. In addition, they had the transverse potential to answer to greetings, requests for help, and chit-chat questions asked to chatbots. BaSCo is a compendium of 1377 tagged utterances with every sample annotated at three levels: (i) NLU semantic labels, considering intents and entities, (ii) code-switching proportion, and (iii) domain of origin.

Keywords: Code-Switching, Language-Mixing, Corpus Compilation, Semantic Annotation, Language Proportion, Natural Language Understanding, Dialogue Systems

1. Introduction

The phenomenon of code-switching –i.e., alternating words belonging to two or more different languages or "codes" at utterance boundaries (intersentential code-switch) or within the same utterance (intrasentential code-switch) (Zirker, 2007)– is very common in both oral and written interactions in diglossic areas such as the Basque Country (Barredo, 1997). Basque is an isolated European language that originates from and is spoken mainly in the Basque Country, where currently 32% of its population is bilingual[1].

A common feature of natural interactions among such bilingual speakers is the spontaneous continuous switching between the Basque and Spanish languages. This communicative strategy goes by the popular name of *Euskañol*, which is a blending of the words "Euskara" (Basque) and "Español" (Spanish).

Despite the widespread universality of code-switching, annotated linguistic resources encompassing code alternations are very scarce (Hamed et al., 2017). For the case of Euskañol, since it is a quite narrowly localised phenomenon, the situation is even harsher. To the extent of our knowledge, only a few research papers have been published on this matter: Lantto (2014) discusses the relationship between swearing and slang to Basque-Spanish code-switching. Ezeizabarrena and Aeby (2010) study the syntax of Euskañol based on observations of interaction situations between children and adults. However, both articles approach code-switching from the standpoint of linguistic analysis and do not explore the usage of this data for the development of language technologies like chatbots (i.e., conversational agents). Furthermore, none of the mentioned studies provides a public corpus. The ES-Port corpus (García-Sardiña et al., 2018) contains some cases of annotated Basque-Spanish code-switching and is publicly accessible. However, the main language of the corpus is Spanish, and usage of Euskañol appears only as a by-product.

Nowadays, the most groundbreaking techniques in Natural Language Processing (NLP) tasks are entirely data-driven. This causes the scarcity of annotated data to be a major issue and, as previously mentioned, this lack of data is much greater when it comes to texts involving code-switching. Nowadays Pre-Trained Models (PTM), which are trained on large corpora and capable of understanding universal linguistic representations that are usable for subsequent NLP tasks, often come as a solution to such problem. However, annotated corpora are still needed to adapt these PTM to specific tasks and use cases.

In order to understand the communicative intention of the user's input –and thus be able to act accordingly– chatbots usually employ an intent/entity based annotation schema to train data-driven NLU systems (Chen et al., 2017). These systems, as they are data-dependant, are able to work with the languages which were employed in the training set.

When facing intrasentential code-switching, current algorithms have serious problems understanding key intents and entities, as the same sentence would contain words or phrases belonging to two or more grammatical systems or subsystems (Gumperz, 1982). One possible approach to solving this problem is based on employing a single Natural Language Understanding (NLU) model trained with data written in multiple languages: a multilingual model.

Contextualised multilingual models, such as mBERT (Devlin et al., 2019) and XLM-R (Conneau and Lample, 2019), have achieved state-of-the-art results on monolingual and multilingual tasks in NLU benchmark tests (Wang et al., 2018; Hu et al., 2020; Liu et al., 2020). However, the effectiveness of these models on code-switching tasks remains

[1] *V Encuesta Sociolingüística del Gobierno Vasco*, 2013

unknown (Winata et al., 2021). This is partly due to the lack of resources and annotated data manifesting this phenomenon.

The present work introduces the BaSCo –*Ba*sque-*S*panish *Co*de-Switching– corpus and its public release[2]. This corpus is a compendium of 1377 utterances in Euskañol from the administrative, fiscal, transportation, social, and general domains. All the texts in the corpus have a reference sentence in Spanish or Basque from which they have been derived and are labelled with intents and entities. The reference samples were extracted from the texts used for training the NLU modules of several chatbots of the above-mentioned domains.

The present article is structured as follows: Section 2. discusses (i) the origin and main characteristics of the source data and (ii) the detailed procedures that have been followed to create the final Euskañol corpus. Section 3. presents the pre-processing of the corpus and the criteria that determine the validation and annotation of each utterance. In Section 4., the statistics of the final corpus are provided. Finally, in Section 5., the main contributions of this work are highlighted, enticing use cases of the released corpus are proposed, and tentative lines to develop in a future work are defined.

2. Source Data and Compilation

This section describes the main characteristics of the source data and the compilation process of the *BaSCo* corpus.

2.1. Initial Setup

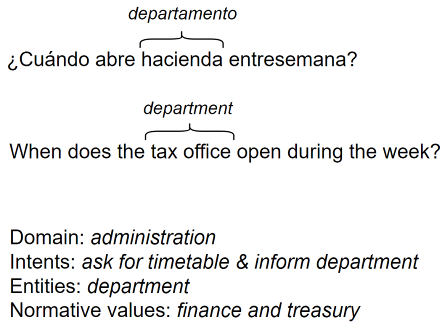
The departure point of this work was a collection of text samples used for training the NLU modules of four task-oriented bilingual chatbots. These were proprietary chatbots developed by the authors in previous projects and were meant to answer specific questions associated with the administration, fiscal, and transport domains. A couple of examples of such utterances include *Nire datu fiskalak internet bidez eska ditzaket?* ("Can I request my tax information online?") and *Para asuntos de tráfico, a dónde tengo que dirigirme?* ("For traffic matters, where should I turn to?") for the fiscal and transport domains respectively.

Therefore, the sentences comprising the initial corpus were classified into five categories depending on their domain: the `administration` domain, the `transport` domain, the `fiscal` domain, the `generic` domain –for transversal intents like greetings, help, assertion, etc.–, and the `social` domain –including intents constituting the social layer of some chatbots, like criticising the chatbot or asking their age–.

In addition to their domain label, the source samples were annotated with semantic information regarding their intents and entity values. Intents are labels that designate the high-level communicative purpose (e.g. to inform, to deny, etc.) of the utterances and are usually extracted by analysing the whole sentence. On the other hand, entities are key elements in the utterance that provide specific information. In some cases, entities can have a set of fixed values associated to them, which are referred to as *normative values*. For example, the normative values of the entity "department" could be "department of health", "department of education" or "department of finance and treasury", as is the case in the example in Figure 1.

Thus, intents evidence the high-level communicative intention, which can be applied on entities with their values. An utterance may have one or more intents, while entities may or may not appear.

Figure 1 shows an annotated example from the source dataset, where the concept "tax office" is an entity classified as "department" and the user's communicative intention, "inform department" and "ask for timetable", are the intents.



departamento
¿Cuándo abre hacienda entresemana?
department
When does the tax office open during the week?
Domain: administration
Intents: ask for timetable & inform department
Entities: department
Normative values: finance and treasury

Figure 1: Annotated utterance example from the source dataset, and its translation to English.

After eliminating repeated samples, as the `generic` and `social` layers were overlapped across chatbots, we ended up with a total of 1936 reference sentences in Spanish and 2216 reference sentences in Basque. Figure 2 shows the number of initial samples in each language per domain in which the strong parallelism between the source data available in Basque and in Spanish can be appreciated.

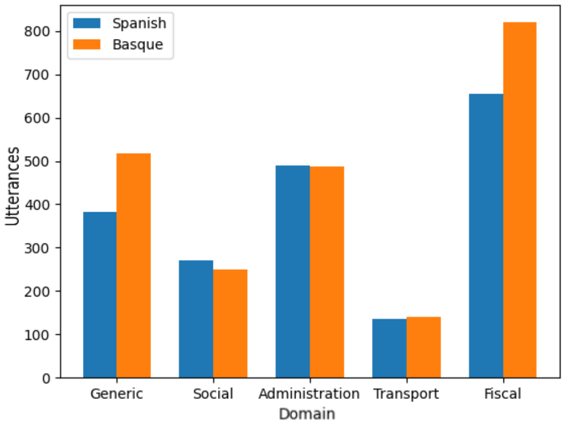


Figure 2: Number of unique samples per domain in the base corpus.

[2] Publicly available under a Creative Commons Share-alike license at `https://github.com/Vicomtech/BaSCo-Corpus`

The number of unique intents and entities associated with each domain is reported in Table 1. These categories are the same in both Basque and Spanish. Intents belonging to more than one domain have been added in all the corresponding rows of Table 1.

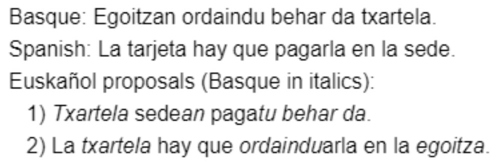
Domain	**Number of Intents**	**Number of Entities**
Generic	9	0
Social	18	0
Administration	36	7
Transport	6	2
Fiscal	26	8
Total	90	17

Table 1: Number of unique Intents and Entities per domain and total sum in the source corpus.

2.2. Code-Switching Data Gathering

A web user interface was created to gather the corpus in Euskañol. This interface displayed a randomly selected reference sentence –which could be in Basque or Spanish– from the source corpus, and an input field for the user to provide an alternative text in Euskañol. There was an option to skip to the next in case the user could not come up with a realistic alternative in Euskañol.

The user interface was distributed among volunteering people in our organisation with some knowledge of both languages. They were instructed to provide an alternative in Euskañol to the provided reference text, which be written in either Spanish or Basque, while keeping its meaning. Figure 3 shows a couple of illustrative examples provided along with the instructions.



Basque: Egoitzan ordaindu behar da txartela.
Spanish: La tarjeta hay que pagarla en la sede.
Euskañol proposals (Basque in italics):
1) *Txartela* sede*an* pagat*u behar da*.
2) La *txartela* hay que *ordaindu*arla en la *egoitza*.

Figure 3: Provided example with references in Basque and Spanish and a proposal in Euskañol (English: "The card must be paid for at the headquarters").

3. Data Curation and Annotation

After the data gathering phase was completed, the compiled raw data needed to be curated and annotated to produce a final corpus.

Three human annotators, all of them bilingual in Spanish and Basque, carried out these tasks using a web user interface which showed them a reference text, the Euskañol proposed version to be evaluated, and the options to validate and annotate it. It should be noted that each of the annotators originates from a different region of the Basque Country where distinct dialects of Basque are spoken.

3.1. Data Curation

The curation phase involved removing duplicates and filtering out which utterances were or were not valid for the target Euskañol corpus.

To do so, a set of guidelines to determine the validity of an utterance was established. An utterance would be considered valid if:

- It is compliant with the task objective: the utterance is, to whatsoever extent, in a mixture of Spanish and Basque.

- From a semantic point of view, its content remains the same as its reference text's: the same NLU labels are valid for both the reference utterance and the new one in Euskañol. Small differences in the utterance's tone or slightly different nuances are overlooked if the general meaning is preserved.

- It sounds natural: it could be an utterance that a person would use in a real conversation in Euskañol, it does not sound artificial.

Following these guidelines, only those utterances that were considered valid by at least 2/3 of the annotators would be included and annotated in the final corpus[3].

3.2. Corpus Annotation

After invalid samples were filtered out, the next phase involved the annotation of the valid utterances. This corpus is annotated at three levels: (i) NLU semantic labels, (ii) code-switching proportion, and (iii) domain of origin of the source chatbot. Each of these is further explained below.

3.2.1. NLU Annotation

As explained above, the source data was made of text samples used for training the NLU modules of several task-oriented bilingual chatbots. Such source samples were semantically annotated with intents, entities, and their normative values when needed.

Given that the new Euskañol utterances were linked to an annotated reference and that one of the filtering criteria was that the meaning was kept, the semantic labels regarding the intents and entities of the new sample should remain the same as their reference text's. Entities' values, on the other hand, needed to be adapted to the new utterance, given that the textual values and their positions in the sample could have probably changed when producing the code-switched utterance.

The example in Figure 4 shows an utterance in Euskañol at the bottom and its linked source sample in Spanish above. As can be seen, although the type of entity included remains the same (*document*), its value in the text and its indices have changed.

[3] The discarded samples are also supplied –fully annotated– together with the final corpus, in case others would consider them of interest.

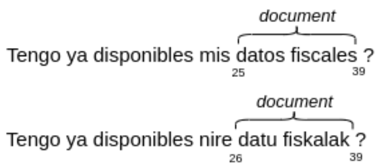

document
Tengo ya disponibles mis datos fiscales ?
25
39
document
Tengo ya disponibles nire datu fiskalak ?
26
39

Figure 4: Example of an entity-annotated reference sample and an Euskañol version proposed for it (English: "Do I already have my *tax data* available?").

3.2.2. Code-Switching Level Annotation

An additional annotation level includes annotator's perspectives on the proportion of Basque and Spanish constituting the new Euskañol utterance. Three main classes were defined for this level:

- `more-es` label: if it is considered that the utterance includes a larger proportion of Spanish than Basque. E.g. "*orduan* lo recibirá mi novia?" ("then my girlfriend will receive it?", Basque in italics).
- `more-eu` label: if it is considered that the utterance includes a larger proportion of Basque than Spanish. E.g. "nola eskatzen da *cita*?" ("how do I make an appointment?", Spanish in italics).
- `balanced` label: if it is considered that the proportion of Basque and Spanish is more or less balanced. E.g. "ayuda *mesedez*" ("help please", Basque in italics).

Annotators' considerations on the language proportion were not necessarily based on number of words or even morphemes in each utterance, but it was rather a more general perspective on the whole sample based on native speaker intuition. Going deeper into this matter is not the main goal of the current corpus.

In cases where annotators reached a tie between the three possible labels, the final label for the sample would be `balanced`.

3.2.3. Domain of Origin

Given that the source data was obtained from task-oriented chatbots, a label referring to the source domain was automatically included. As stated in Subsection 2.1., the possible domain labels are: `administration`, `transport`, `fiscal`, `generic`, or `social`.

4. Corpus Statistics

In the Euskañol data gathering phase, a total of 1656 sentences were collected via the web interface. After an initial filtering process to eliminate repeated samples, a total of 1606 utterances remained.

Next, the annotators filtered which of these 1606 sentences were valid and which were not according to the criteria defined in Section 3.1.. Simultaneously, they ranked the proportion of Euskañol in each utterance. As a result, a total of 1377 examples were filtered as valid and 229 were discarded.

Note that from now on whenever the *final corpus* is mentioned it will refer exclusively to these 1377 samples in Euskañol.

The Inter-Annotator Agreement (IAA) was calculated by using the Fleiss' kappa, and the result of determining which utterances were valid and which were not was $\kappa = 0.4998$, which translates into a moderate agreement (Viera et al., 2005).

The IAA on the language proportion label was calculated only on the valid samples and resulted in $\kappa = 0.6083$, which translates into a substantial agreement. This score could be due to the dialectal differences between the annotators, as it affects on how the *Euskañolised* utterances are formed and what may sound "natural" to them.

The snippet below shows the final JSON data structure using a real example from the corpus.

```
"referent": "dónde está la casa del
↪  deporte?",
"source_lang": "es",
"domain": "administration",
"intents": [
  "preguntar|ubicacion",
  "informar|tipo-oficina"
],
"entities": [
  {
    "entity": "tipo-oficina",
    "value": "casa del deporte",
    "normative_value": "deportes",
    "start": 14,
    "end": 29,
    "type": "bounded"
  }
],
"code_switching": [
  {
    "text": "Casa de deporte non dago?",
    "entities": [
      {
        "entity": "tipo-oficina",
        "value": "Casa de deporte",
        "normative_value": "deportes",
        "start": 0,
        "end": 14,
        "type": "bounded"
      }
    ],
    "lang_proportion": "balanced"
  }
]
```

Each reference sample –key "referent"– can include one or more Euskañol proposals –key "code-switching"– that derive from it. The minimum number of labeled intents for a sample in the resulting corpus was 1, and the maximum was 4. Likewise, the number of entities varies between 0 and 4.

The "value" of an entity in "entities" designates the text span in that particular statement that corresponds to the entity. The "normative value" represents the standardised equivalent of the text value. The "type" key can have two possible values: *bounded* or *free*. If the type is "bounded", it means that there is a limited set of normative values that the entity can have. On the other hand, if the type is “free”, the possible normative values that the entity can take are infinite. An example of the latter case would be the entity *person-name*, whose normative values would not be predefined since the possibilities are uncountable. Finally, the “start” and “end” keys simply denote the starting and ending indices of the span in the text corresponding to the entity.

The total number of samples per domain in the BaSCo corpus is reported in Table 2. Note that not all source utterances were provided with valid code-switched proposals (e.g. one- or few-word expressions like *agur*, “bye”), resulting in a final corpus with a smaller number of samples than the source corpus.

Domain	**Basque**	**Spanish**	**Euskañol**
Generic	517	383	119
Social	250	271	205
Administration	488	490	538
Transport	141	136	55
Fiscal	820	656	460
Total	2216	1936	1377

Table 2: Number of samples per domain and total in the source and final corpora.

The number of unique intents and entities that appear in the final corpus and their total number of occurrences are presented in Table 3. If this table is compared with Table 1 it can be observed that two intents are missing, one belonging to the `administration` domain and the other one to the `fiscal` domain. The BaSCo corpus covers the remaining intents and all entities of the original corpora. The number of times these intents and entities are labeled can also be consulted in Table 3.

Domain	**Intents**	**Entities**
	(Unique/Freq.)	(Unique/Freq.)
Generic	9 / 191	0
Social	18 / 205	0
Administration	36 / 956	7 / 389
Transport	6 / 55	2 / 9
Fiscal	26 / 924	11 / 552
Total	90 / 2331	20 / 950

Table 3: Number of unique and total occurrences of Intents and Entities per domain and total sum in the final corpus.

The average number of samples in Euskañol labeled with each intent is 26.19. The most repeated intent in the final corpus has a total of 246 occurrences while the least repeated one appears only twice. On the other hand, the average number of samples for each entity is 52.6. The most repeated entity in the corpus appears 202 times and the least repeated one presents five occurrences.

As regards language proportion, Figure 5 shows the number of Euskañol samples per label and language of the reference text. As it can be seen, the Basque language is predominant in the code-switching samples, even if the source sample is written in Spanish.

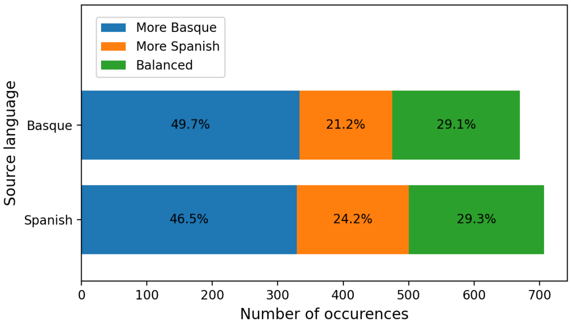


Figure 5: Language proportion statistics in the final corpus considering language of the source text.

In the final corpus there are 313 utterances tagged as `more-es`, 662 as `more-eu` and 402 as `balanced`. The reference samples of these utterances were 707 with Spanish as source language and 670 with Basque as source language.

Finally, the average number of words per utterance in the corpus is 5.43, the 49.2% of the Euskañol samples being between 3 and 5 words in length and the maximum length of a text being 21 words.

5. Conclusions

In this work the very first chatbot-oriented Euskañol corpus has been compiled and made publicly available. It contains 1377 utterances in Euskañol, revised, filtered, and labelled by three bilingual annotators. These labels correspond (i) to the semantic level of the samples, being the multiple intents and entities of each text labelled, (ii) to the domain to which the utterance belongs (e.g. administration, transport...), and (iii) to the overall proportion of Basque and Spanish contained in each sample.

Thus, the means to start overcoming the linguistic reality of code-switching between Basque and Spanish in spoken and written spontaneous language interactions has been provided. In this way, progress has been made towards building adaptive, multilingual conversational assistants that do incorporate code-switching strategies and can therefore communicate fluently with multilingual users.

The BaSCo Corpus can be potentially useful for tasks like the following:

- **Multilingual chatbots.** Compare the capacity of different language representation models when it comes to understanding intents and entities in the case of a corpus containing Basque-Spanish code-switching.

- **Speech recognition.** Develop speech to text systems that can perform adequately when the input audio source is given in Euskañol. Previous work has shown that language models that include a code-switching prediction module outperform previous baselines when it comes to mixed language speech recognition tasks (Li and Fung, 2012; Li and Fung, 2013).

- **Linguistic analysis.** Some of the linguistic studies that can be carried out with this corpus are: exploring the most common structures and characteristics of Euskañol and thus extracting the different patterns of its construction; taking the first steps towards obtaining a formal grammar of Basque-Spanish code-switching; quantifying and visualising the nature of the integration of both languages in code switching, etc.

- **Performance evaluator.** So far, there are hardly any resources available to evaluate how much services –like dialogue systems or speech recognisers– are degraded by the phenomenon of code-switching. This corpus can be used to assess the quality of these services in the domains covered.

- **Language identifier evaluator**. Evaluate the performance of language detectors by using the *proportion* labels and analyse their behaviour in cases where the label is tagged as *balanced*.

Currently, the BaSCo corpus is being used to test the performance of three multilingual BERT (mBERT) (Devlin et al., 2019) models on their learning abilities in bilingual Basque and Spanish chatbots. Future work includes increasing of the size of the corpus in order to create a richer resource or incorporating additional annotation levels so that it can be used in other applications, such as sentiment analysis.

6. Acknowledgements

The authors of this work would like to thank the volunteers who participated in the compilation phase providing their Euskañol proposals.
This project has received funding from the the Department of Economic Development and Infrastructure of the Basque Government under grant number KK-2020/00055 (EKIN).

7. Bibliographical References

Barredo, I. M. (1997). Pragmatic functions of code-switching among basque-spanish bilinguals. *Retrieved on October*, 26:2011.

Chen, H., Liu, X., Yin, D., and Tang, J. (2017). A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, 19(2):25–35.

Conneau, A. and Lample, G. (2019). Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems*, 32:7059–7069.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Ezeizabarrena, M.-J. and Aeby, S. (2010). Les phénomènes de code-switching dans les conversations adulte-enfant (s) en basque-espagnol: une approche syntaxique. *Corpus*, (9).

García-Sardiña, L., Serras, M., and Del Pozo, A. (2018). Es-port: a spontaneous spoken human-human technical support corpus for dialogue research in spanish. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Gumperz, J. J. (1982). *Discourse strategies*. Cambridge University Press.

Hamed, I., Elmahdy, M., and Abdennadher, S. (2017). Building a first language model for code-switch arabic-english. *Procedia Computer Science*, 117:208–216.

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., and Johnson, M. (2020). Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Lantto, H. (2014). Code-switching, swearing and slang: The colloquial register of basque in greater bilbao. *International Journal of Bilingualism*, 18(6):633–648.

Li, Y. and Fung, P. (2012). Code-switch language model with inversion constraints for mixed language speech recognition. In *Proceedings of COLING 2012*, pages 1671–1680.

Li, Y. and Fung, P. (2013). Improved mixed language speech recognition using asymmetric acoustic model and language model with code-switch inversion constraints. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7368–7372. IEEE.

Liu, Z., Winata, G. I., Lin, Z., Xu, P., and Fung, P. (2020). Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8433–8440.

Viera, A. J., Garrett, J. M., et al. (2005). Understanding interobserver agreement: the kappa statistic. *Fam med*, 37(5):360–363.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Winata, G. I., Cahyawijaya, S., Liu, Z., Lin, Z., Madotto, A., and Fung, P. (2021). Are multilingual models effective in code-switching? In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 142–153.

Zirker, K. A. H. (2007). *Intrasentential vs. intersentential code switching in early and late bilinguals*. Brigham Young University.