How good are humans at Native Language Identification? A case study on Italian L2 writings

Elisa Di Nuovo
Elisa Corino
Dipartimento di Lingue e Letterature
Straniere e Culture Moderne
University of Turin
elisa.{dinuovo,corino}@unito.it

Cristina Bosco
Dipartimento di Informatica
University of Turin
bosco@di.unito.it

Abstract

In this paper we present a pilot study on human performance for the Native Language Identification task. We performed two tests aimed at exploring the human baseline for the task in which test takers had to identify the writers' L1 relying only on scripts written in Italian by English, French, German and Spanish native speakers. Then, we conducted an error analysis considering the language background of both test takers and text writers.

1 Introduction

Native Language Identification (NLI) is a task usually performed by machines consisting in identifying the mother tongue (henceforth L1) of a person based only on their writings in another language (e.g. L2 or L3[1]). To date, the majority of the existing studies have focused on English as L2, that is English used by people who are acquiring English as a second or foreign language (Tomokiyo and Jones, 2001; Koppel et al., 2005; Malmasi et al., 2015; Kulmizev et al., 2017; Markov et al., 2017; Cimino and Dell'Orletta, 2017, among others). Three editions of the NLI shared task had been organized (Tetreault et al., 2013; Schuller et al., 2016; Malmasi et al., 2017) in which systems had to correctly identify the L1 among 11 L1s.

The basic assumption of this task is that when we learn a new language (henceforth Target Language, TL), our L1 interferes within the learning process introducing in the TL productions clues that can be automatically detected. Nevertheless, another issue to be investigated within this task is the interference in the TL learner's productions also of other languages previously learned as L2 or L3. In fact, L1 may not be the only language playing a role in the acquisition of a TL, since "bi- or multilingualism is as frequent in the population of the world as pure monolingualism, perhaps even more frequent" (Hammarberg, 2001, p. 21). This issue is especially relevant performing the NLI task in languages other than English. For instance, when someone learns Italian, it is likely their L3, since English is the language taught worldwide as L2 (with more that 100 million learners, as stated in the British Council annual report 2018-19).

In this paper, we investigate the human performance for NLI applied on productions of learners Italian, thus focusing not only on the issues related to *second language acquisition* (Ellis, 2015; Slabakova, 2016), but also to *third language acquisition* (Cenoz et al., 2001; Picoral, 2020). We asked human Test Takers (TTs) to perform a simplified NLI task on learner Italian scripts extracted from VALICO (Corino and Marello, 2017), selecting only four L1s[2]—i.e. English (EN), French (FR), German (DE) and Spanish (ES). This simplified task will be challenging since all the selected languages are Indo-European languages sharing typological similarities. Moreover, we performed an error analysis of the test results considering learners' L1 and L2(s) and learning stage (i.e. year of study), in addition to text features and TTs' language background. Test results could be useful for the improvement of the error annotation scheme introduced in Di Nuovo et al. (2019).
Our research questions therefore are:
1. How good are humans in performing the NLI task?
2. What features share the most problematic texts?
We try to answer to these questions in this paper organized as follows: in Section 2 we briefly de-

[1] L3 is any language acquired in addition to someone's L1 and another or more L2s.

[2] In evaluation campaigns, NLI systems are trained and tested on 11 L1s.

scribe previous work on the subject; in Section 3 we describe the tests performed and discuss the results; in Section 3.2 we conduct the error analysis; and in Section 4 we conclude the paper.

2 Related work

To the best of our knowledge, the first study assessing the human ability in identifying members of the same native language group is that of Ioup (1984). She demonstrated that native speakers were able to identify these groups only when relying on phonological cues, supporting the assumption that "syntactic errors in L2 acquisition cannot be accounted for as a result of the transfer of L1 patterns" (Broselow and Newmeyer, 1988, p. 197).

Odlin (1996) proved instead that readers who know the observed L1s (Korean and Spanish) can distinguish certain syntactic patterns in L2 English texts. Nowadays, his hypothesis is supported by the improved accuracy of machine learning systems using syntactic information (Malmasi et al., 2013; Markov et al., 2017, among others), such as PoS tags, Context-free Grammar Production Rules, Tree Substitution Grammar fragments, Grammatical Dependencies (see Malmasi (2016) for more details).

A more recent study simplified the NLI task to be performed by humans (Malmasi et al., 2015). The authors selected from TOEFL 11 (Blanchard et al., 2013)[3] 30 texts (6 per each of the 5 languages included: Arabic, Chinese, German, Hindi, and Spanish), 15 considered *easy* and 15 *hard*, according to the ease with which most systems predicted the L1. They chose ten raters, all professors or researchers who might have had exposure to the 5 selected L1s. The average accuracy achieved by raters (37.3%, top rater 53.3% and lowest 30%) shows how difficult the task can be for humans. Their approach does not give attention to text features and writer's language background that, in addition to experts' knowledge of the 5 L1s involved, could have had an impact on the task performance.

Two NLI experiments on English L2 performed by humans are also reported by Jarvis et al. (2019). In both cases, humans were asked to identify in L2 English texts writers of their same L1. In the first experiment, six Finnish speakers were asked to identify the author of the L2 English text as being a Finnish or Swedish native speaker. In the second experiment, in addition to the six Finnish raters, participated ten Spanish-speaker raters and all had to identify if the writers' L1 shared their same L1 (Spanish or Finnish). The features that lead the raters to their decision were used to discuss the results achieved (in the second experiment, over 80% accuracy for the top Finnish raters and over 90% for the top Spanish raters). It is important to note that the accuracy achieved in Malmasi et al. (2015) and Jarvis et al. (2019) is not comparable since the experiment settings are completely different.

In this paper we describe experiments which are more similar to that carried out by Malmasi et al. (2015). However, differently from the papers described, we focus on data extracted from the VALICO corpus and in Italian, a language for which there are no previous studies about human NLI (see Malmasi and Dras (2017) for a multilingual approach to NLI in which they developed also a system for L2 Italian, using precisely texts collected from VALICO).

3 Test Description

We asked our TTs to perform two tests. The first (Test 1) is a simplified 4-class NLI test and the second (Test 2) a sort of guided Logistic Regression.

To Test 1 participated 25 TTs, 11 of them to Test 2.[4] They are all Italian native speakers and were selected according to their knowledge of Italian as L2, foreign languages or linguistics. Our average TT has a master's degree in a field related to Linguistics or Foreign Languages and speak on average two additional languages among the selected L1s.

The selected L1s—FR, DE, ES and EN—are the most commonly taught in Italy, so theoretically it would have been easier to find human experts knowing them[5]. In addition, these four languages represent two different families: while FR and ES belong to Romance languages like Italian, EN and DE belong to Germanic languages. For this reason we believe the task, although simplified (because constrained to only four languages), challenging

[3] TOEFL 11 is the dataset used in NLI shared tasks.

[4] We would like to stress the difficulty we faced in finding the TTs, not only due to the skills required, but also to the time and concentration required to perform these not at all short tests. We want also to point out that Test 1 is the experiment, to date, featuring the highest number of TTs, which are 10 in Malmasi et al. (2015) and 16 in Jarvis et al. (2019).

[5] However, we had difficulties in finding human experts knowing all the four languages. Only three out of twenty five know all the four considered languages.

enough. Furthermore, we expect different transfer patterns from the speakers of the two families.

Test 1 is a simplified NLI task, namely a multiple-choice test made of 48 questions. Each question contains a short text written in Italian by a non-native speaker and the TT had to identify the writer's L1 choosing it between EN, FR, DE and ES. Texts were randomly selected from VALICO-UD (Di Nuovo et al., 2019) with an almost balanced distribution with respect to the L1: 11 EN, 14 FR, 10 DE and 13 ES—TTs were not aware about this distribution. Their length ranged from 58 to 484 words (mean length 136.19 words, standard deviation (SD) of 64.73 words). There were no statistically significant differences in length between the L1 groups as determined by one-way ANOVA ($F = 2.24$, $p = .09$).

Test 2 consisted of 24 texts drawn from Test 1 according to the difficulty human TTs had in identifying the correct L1. The TTs were asked to assign a percentage to each of the four L1s involved, performing a sort of guided Logistic Regression: the higher the percentage assigned to a language, the higher chance of being the writer's L1.

Table 1 resumes the information about the number of texts written by EN, FR, DE, ES native speakers (# Text/L1), and the percentage of TTs knowing that L1 (TT/L1) for both Test 1 and 2.

—	L1	# Text/L1	TT/L1
Test 1	EN	11	100%
	FR	14	72%
	DE	10	28%
	ES	13	56%
Test 2	EN	5	100%
	FR	10	82%
	DE	7	36%
	ES	2	73%

Table 1: Test 1 (25 TTs) and 2 (11 TTs) in figures.

3.1 Test Results and Discussion

The best result on Test 1 was achieved by TT1, correctly identifying the L1 of 26 out of 48 texts (54.2% accuracy), the lowest by TT25, correctly predicting the L1 of only 10 texts (20.8% accuracy). TT1 and TT25 are both PhD students in Digital Humanities: TT1 speaks EN and ES, while TT25 EN and DE. Table 2 shows score, accuracy (Acc.) and TTs' teaching experience (TTE)[6].

[6]As teachers of Italian as L2, EN, ES, FR or DE

These results suggest that, in our sample, teaching experience as well as knowing all the four L1s involved (as TT2, TT15 and TT23 did) were not discriminant factors. In addition, we checked if TTs speaking EN, FR, DE or ES could improve the identification of that L1(s) they know, but we did not find significant difference. Conversely to what stated by Jarvis et al. (2019, p. 222), TT18—knowing FR and Italian as L1 and EN as L2—did not identify more FR texts than other TTs knowing FR as L2.

TT	Score	Acc.	TTE
TT1	26	54.2%	yes
TT2, TT3, TT4, TT5	25	52.1%	3 yes, 1 no
TT6, TT7, TT8	24	50.0%	2 yes, 1 no
TT9	23	47.9%	no
TT10, TT11, TT12	22	45.8%	3 yes
TT13, TT14, TT15	21	43.7%	1 yes, 2 no
TT16	20	41.7%	yes
TT17	19	39.6%	yes
TT18	18	37.5%	yes
TT19	17	35.4%	yes
TT20	15	31.2%	yes
TT21, TT22, TT23, TT24	13	27.1%	1 yes, 3 no
TT25	10	20.8%	yes
Mean	20	**SD** (σ)	4.7

Table 2: Test 1 results - TT, score and accuracy achieved and TTE.

We classified the texts into 2 major categories, *easy* and *hard*, according to the percentage above or below 50% of correct answers assigned by the TTs, respectively. In total we have 17 easy texts (10 ES, 3 FR, 3 EN, 1 DE) and 31 hard texts (3 ES, 11 FR, 8 EN, 9 DE). Almost the totality of ES texts were identified by more than 50% of TTs, but we will comment on this in the next sections.

We further divided these two categories into subcategories. Easy texts are divided into *Clear-cut* texts, in which authors' L1 has been easily identified by almost all the TTs, and *Confusing* texts, identified by the majority of the TTs but a number leaned towards the same wrong class. Hard texts are further divided as *Scattered*, *Wrong Scattered*, and *Wrong Clear-cut*. In *Scattered* texts, votes are spread across two or three L1s, but the L1 receiving the majority of votes is the correct one. In *Wrong Scattered*, votes are spread across two or three L1s, but this time, the L1 receiving the majority of votes is a wrong one. Finally, in *Wrong Clear-cut* texts the L1 receiving more than 50% of the votes is an incorrect one.

Figure 1 shows the texts divided into these five categories. In the x axis we have the four possible

L1s and in *y* axis the text identification number. It is interesting to notice the similarity of Clear-cut and Wrong Clear-cut categories. In both categories, more than 50% of TTs opted for the same L1. The only difference relies in the fact that in the former it is the correct prediction, in the latter a wrong prediction.

Our hypothesis is that texts in the same subcategory share similar features. All Clear-cut texts contain spelling errors, literal translations or calques that clearly and explicitly lead to one of the four L1s (e.g. *cuando* for ES, *bizarre* for FR, *piance* for DE). Confusing texts had some ambiguous clues (e.g. *qualquosa* which can be written by ES or FR speakers, but for different reasons) which may cast doubt on at least 2 L1s. The main clues in Scattered texts were at a grammatical level (e.g. -ing form used instead of relative clauses, wrong agreement in gender and number), so TTs had to pay more attention. In Wrong Scattered and Wrong Clear-cut texts there were shallow clues, such as misspellings and loanwords—as in Clear-cut texts—(for example *basura*, ES word for 'garbage') which might be indicative of negative transfer, but—differently from Clear-cut texts—these clues were not due to the L1 (in our example EN) but to other known L2s (in this case ES). However, in most of the cases, L2 transfers were in conjunction with L1 clues (such as the use of *ritornare* instead of *restituire* in "rotornare la borsa a Maria", literal for 'to return the bag to Maria', or also *nel piano* instead of *per terra*, probably a translation error due to the polysemy of 'floor') that our TTs did not notice or considered less relevant clues. In order to clarify TTs recognition of the clues, we created another test, featuring the same 48 texts, but this time we told the TTs to highlight the clues that they used to make their prediction. We cannot provide information about this because we are still collecting the answers.

However, besides the clues, we wanted to capture also TTs' uncertainty. Thus, in Test 2, we asked our TTs to assign a percentage to each L1 of the 24 most challenging texts. Figure 2 shows the average results for each L1 per text (correct answer in bold) divided into the subcategories of hard texts: Scattered (S), Wrong Scattered (WS), Wrong Clear-Cut (WCC).

Broadly speaking, there was high uncertainty among the TTs, not always in line with our hypotheses. Even when TTs were particularly sure about one of the four L1s (e.g. assigning 99% to a L1 and 0% to the others), it was not always the correct L1, nor was explainable by writer's L2 knowledge. For example in T14, T21, T48, most of TTs thought that ES was the correct L1. Although these texts showed numerous ES-like transfers (also syntactic ones, e.g. ES personal 'a' in *aiutare **a** la donna*, meaning 'to help the woman'), the writers were FR native speakers, and only one of them claimed to know ES as L2. Still, in the majority of texts, the correct L1 received one of the two highest percentages, suggesting that L1 cues are present and that the TTs correctly interpreted them. It is also interesting to notice that also the analysis conducted by Malmasi (2016, p. 84) about the Oracle classifier suggests that the correct L1 is often in the top two predictions.

3.2 Error Analysis

We calculated recall, precision, accuracy and F1 aggregating all the TTs' Test 1 answers per class; results—in terms of precision (Pre.), recall (Rec.), accuracy (Acc.) and F1—are shown in Table 3. As known, accuracy is influenced by the class distribution, hence F1 is a better metric in this case.

Overall, it was a challenging task as expected. On the one hand, ES proved to be easier to identify than the other three L1s (F1 score 56% compared to FR 38%). However, since all texts belonged to different proficiency levels not balanced across L1s, we cannot say if it is due to easily recognizable L1 patterns or to different interlanguage stages. On the other hand, DE proved to be the hardest L1 to identify. This could be motivated by the fact that only 28% of the TTs that participated to Test 1 have studied DE as L2. It is interesting to notice that DE is the most confused L1 to identify also in the experiment carried out by Malmasi et al. (2015) as clearly stated in Malmasi's PhD thesis (2016, p. 88).

Class	Pre.	Rec.	Acc.	F1
EN	0.34	0.38	0.69	0.36
FR	0.41	0.36	0.66	0.38
DE	0.35	0.30	0.74	0.32
ES	0.52	0.60	0.74	0.56

Table 3: Test 1 aggregated results per class.

Linking these results with TTs' language knowledge and writers' L2 background, we can speculate that TTs' language knowledge in itself

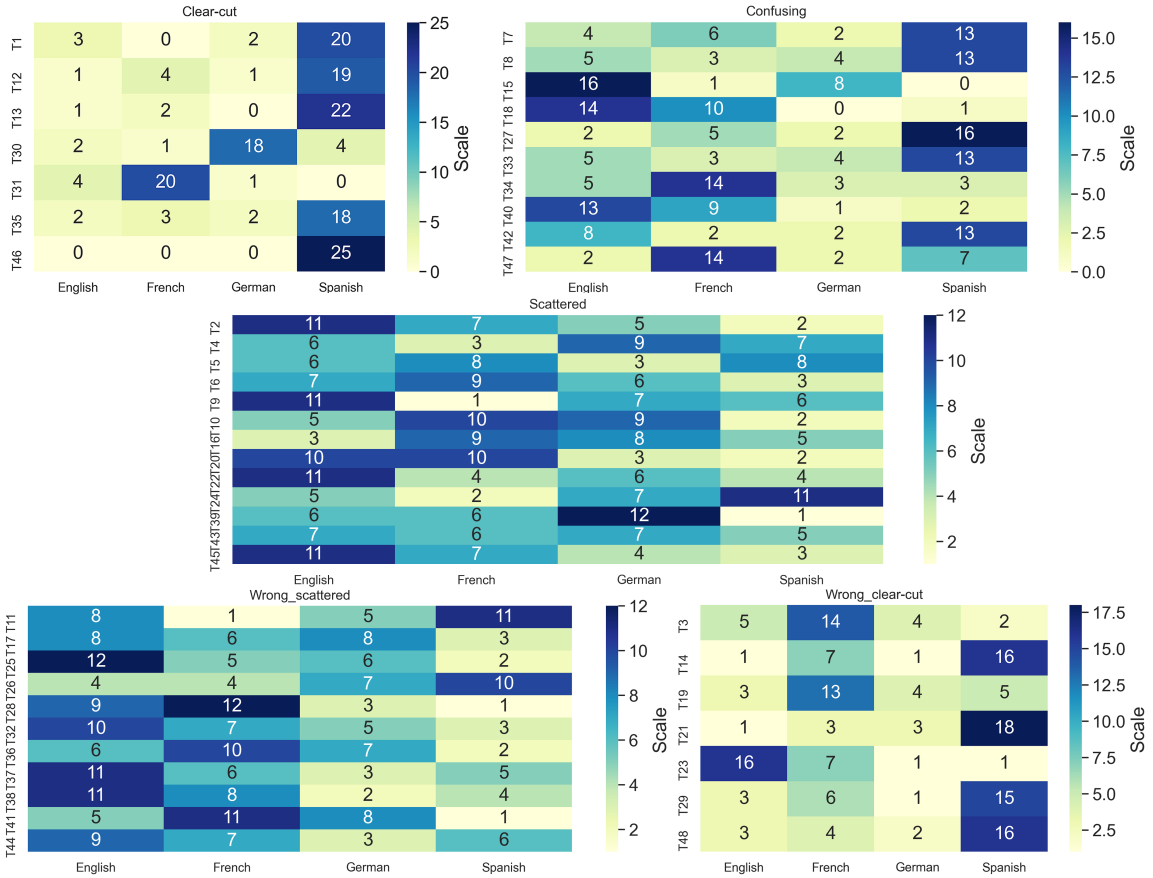

Figure 1: The five text categories stemmed from Test 1.

is not enough to improve L1 identification. In fact, all the TTs know EN as L2, but ES was the language identified correctly most of the time. This might be due to the fact that also all of the writers (except EN native speakers) know EN as L2, making the identification of EN native speakers harder.

We then plotted a confusion matrix (reported in Figure 3) to see the trends per class. Surprisingly, EN, correctly identified 38% of the time, was almost equally confused with the other three L1s (slightly more with FR and ES, 22% and 20% respectively, than with DE, 19%). FR, correctly identified 36% of the time, was confused more with ES (25%) and EN (24%) than with DE (15%). DE, correctly identified 30% of the time, was rarely confused with ES (13%), but frequently mistaken for FR (29%) and EN (28%). ES, when incorrectly identified (40% of the time), was confused slightly more with EN and FR (15% both) than with DE (10%). This might suggest that there is not a clear distinction between the two language families. In addition, it is interesting to notice that the directionality of the confusion is not always bidirectional for the four L1s (e.g. DE is confused with FR and ES but not vice versa).

4 Conclusion

In this paper we described two human NLI tests for Italian. Although it was a simplified NLI task, tailored bearing in mind human skills, it proved to be a difficult task even for experts.

The error analysis showed that ES was the easiest L1 to identify—correctly identified 60% of the time—while DE the hardest. L2 transfer was misleading, even when L1 clues were present. TTs knowledge of the involved L1s proved not to be a discriminant factor.

It would have been interesting to ask the TTs to point out the clues that supported their answers to be less hypothetical in the discussion, especially when dealing with texts featured by L1 and L2 transfer. For this reason, we asked our TTs to take part in another test based on the same texts in which they have to highlight the clues. At the moment, we are collecting the answers.

In the future, we will test a machine learning system on the same texts to compare its results with those of our TTs.

S	EN	FR	ES	DE
T2	**26%**	33%	27%	27%
T5	31%	**37%**	22%	14%
T6	28%	44%	12%	**19%**
T16	39%	**12%**	16%	42%
T20	35%	**40%**	18%	22%
T43	40%	28%	10%	**27%**

WS	EN	FR	ES	DE
T11	**21%**	10%	47%	37%
T17	22%	**26%**	23%	22%
T25	40%	19%	14%	**25%**
T26	**25%**	23%	29%	20%
T28	38%	18%	8%	**35%**
T32	34%	**16%**	8%	35%
T36	41%	25%	11%	**20%**
T37	54%	18%	17%	**14%**
T38	45%	**25%**	22%	18%
T41	**33%**	17%	26%	30%
T44	31%	25%	**36%**	20%

WCC	EN	FR	ES	DE
T3	40%	44%	13%	**16%**
T14	16%	**17%**	63%	11%
T19	25%	24%	**20%**	21%
T21	14%	**22%**	54%	14%
T23	35%	**34%**	9%	23%
T29	**25%**	8%	45%	25%
T48	20%	**25%**	56%	14%

Figure 2: Test 2 results aggregated for L1 and text.

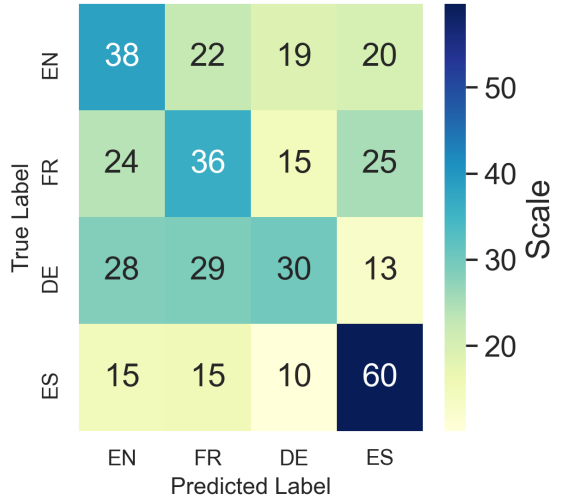

Figure 3: Confusion matrix with Test 1 aggregated data.

Acknowledgements

We would like to thank the anonymous reviewers as well as our TTs, who have dedicated their time to us by carrying out these long tests which require a lot of concentration: E. L. Baratono, C. Borge, C. Borgia, C. Bosco, V. Calabria, S. Cicillini, A. T. Cignarella, C. Conti, V. De Iacovo, G. Esposito, K. Florio, G. Giaccone, A. Giacosa, L. Giannelli, L. Inserra, I. Iubini, A. Marra, M. Pellegrini, S. Peyronel, F. Poletto, S. Racca, A. Rotolo, M. Sanguinetti, E. Truc, and M. C. Zaccone.

References

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. TOEFL11: A corpus of non-native English. *ETS Research Report Series*, 2013(2):i–15.

Ellen Broselow and F Newmeyer. 1988. Second language acquisition. *Linguistics: The Cambridge Survey: Volume 3, Language: Psychological and Biological Aspects*, pages 194–209.

Jasone Cenoz, Britta Hufeisen, and Ulrike Jessner. 2001. *Cross-linguistic influence in third language acquisition: Psycholinguistic perspectives*. Multilingual Matters.

Andrea Cimino and Felice Dell'Orletta. 2017. Stacked Sentence-Document Classifier Approach for Improving Native Language Identification. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 430–437.

Elisa Corino and Carla Marello. 2017. *Italiano di stranieri. I corpora VALICO e VINCA*. Guerra.

Elisa Di Nuovo, Cristina Bosco, Alessandro Mazzei, and Manuela Sanguinetti. 2019. Towards an Italian Learner Treebank in Universal Dependencies. In *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, volume 2481, pages 1–8. CEUR-WS.

Rod Ellis. 2015. *Understanding second language acquisition 2nd Edition-Oxford applied linguistics*. Oxford university press.

Björn Hammarberg. 2001. Roles of L1 and L2 in L3 production and acquisition. In Jasone Cenoz, Britta Hufeisen, and Ulrike Jessner, editors, *Cross-linguistic influence in third language acquisition: Psycholinguistic perspectives*, volume 31, pages 21–41. Multilingual Matters.

Georgette Ioup. 1984. Is there a structural foreign accent? A comparison of syntactic and phonological errors in second language acquisition. *Language Learning*, 34(2):1–15.

Scott Jarvis, Rosa Alonso Alonso, and Scott Crossley. 2019. Native language identification by human judges. In *Cross-Linguistic Influence: From Empirical Evidence to Classroom Practice*, pages 215–231. Springer.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Automatically determining an anonymous author's native language. In *International Conference on Intelligence and Security Informatics*, pages 209–217. Springer.

Artur Kulmizev, Bo Blankers, Johannes Bjerva, Malvina Nissim, Gertjan van Noord, Barbara Plank, and Martijn Wieling. 2017. The power of character n-grams in native language identification. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 382–389.

Shervin Malmasi and Mark Dras. 2017. Multilingual Native Language Identification. *Natural Language Engineering*, 23(2):163–215.

Shervin Malmasi, Sze-Meng Jojo Wong, and Mark Dras. 2013. Nli shared task 2013: Mq submission. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 124–133.

Shervin Malmasi, Joel Tetreault, and Mark Dras. 2015. Oracle and Buman Baselines for Native Language Identification. In *Proceedings of the tenth workshop on innovative use of NLP for building educational applications*, pages 172–178.

Shervin Malmasi, Keelan Evanini, Aoife Cahill, Joel Tetreault, Robert Pugh, Christopher Hamill, Diane Napolitano, and Yao Qian. 2017. A report on the 2017 native language identification shared task. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 62–75.

Shervin Malmasi. 2016. *Native Language Identification: explorations and applications*. Ph.D. thesis, Macquarie University, Sydney, Australia.

Ilia Markov, Lingzhen Chen, Carlo Strapparava, and Grigori Sidorov. 2017. CIC-FBK approach to native language identification. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 374–381.

Terence Odlin. 1996. On the recognition of transfer errors. *Language Awareness*, 5(3-4):166–179.

Adriana Picoral. 2020. *L3 Portuguese by Spanish-English Bilinguals: Copula Construction Use and Acquisition in Corpus Data*. Ph.D. thesis, The University of Arizona.

Björn Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee K. Burgoon, Alice Baird, Aaron Elkins, Yue Zhang, Eduardo Coutinho, and Keelan Evanini. 2016. The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity Native Language. In *Interspeech 2016*, pages 2001–2005.

Roumyana Slabakova. 2016. *Second language acquisition*. Oxford University Press.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 48–57.

Laura M. Tomokiyo and Rosie Jones. 2001. You're Not From 'Round Here, Are You? Naive Bayes Detection of Non-Native Utterances. In *NAACL*.