ISSN 2033-3544

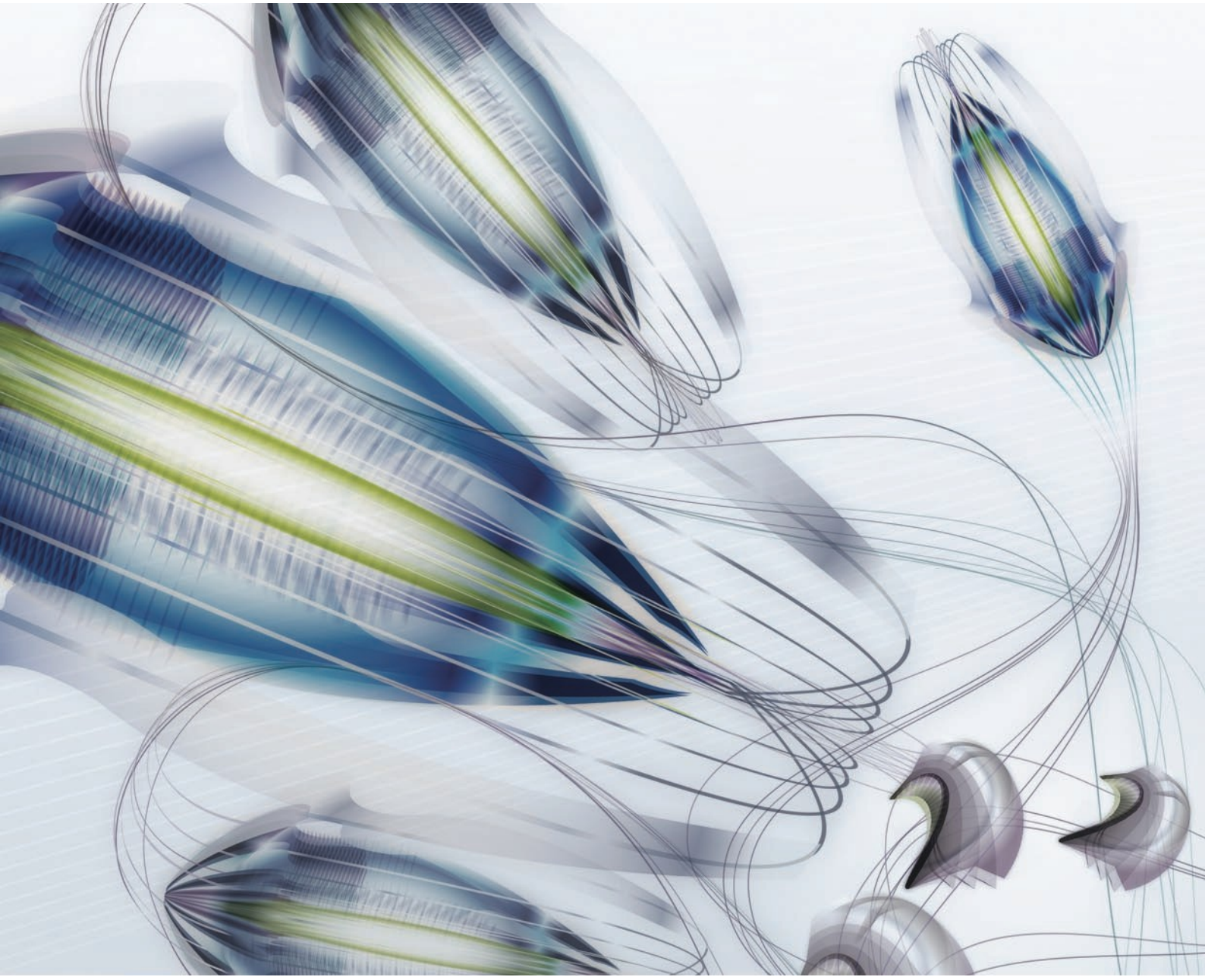
CLiPS 6

Creating TwiSty: Corpus Development and Statistics

Ben Verhoeven, Walter Daelemans & Barbara Plank

COMPUTATIONAL LINGUISTICS & PSYCHOLINGUISTICS
TECHNICAL REPORT SERIES, CTRS-006, MAY 2016

WWW.CLIPS.UANTWERPEN.BE/CTRS



CLiPS
Computational Linguistics & Psycholinguistics
University of Antwerp

Computational Linguistics and Psycholinguistics Research Center
CLiPS Technical Report Series (CTRS)

CTRS-006
May 01, 2016

Creating TwiSty: Corpus Development and Statistics

Ben Verhoeven
CLiPS Research Center, University of Antwerp
Prinsstraat 13, Antwerp, Belgium
ben.verhoeven@uantwerpen.be

Walter Daelemans
CLiPS Research Center, University of Antwerp
Prinsstraat 13, Antwerp, Belgium
walter.daelemans@uantwerpen.be

Barbara Plank
University of Groningen
Groningen, The Netherlands
b.plank@rug.nl

http://www.clips.uantwerpen.be/datasets/twisty

Abstract

This document provides information on the creation of the Twitter Stylometry (TwiSty) corpus (Verhoeven *et al.* , 2016). The corpus contains Twitter profiles annotated with MBTI personality types and gender information, covering six languages: Italian (IT), Dutch (NL), German (DE), Spanish (ES), French (FR), and Portuguese (PT).

- DOWNLOAD: http://www.clips.ua.ac.be/datasets/twisty
- LICENSE: CC-BY-SA 4.0
- ISLRN: 883-383-734-892-8

Reference:
Ben Verhoeven, Walter Daelemans, & Barbara Plank. (2016). TwiSty: a multilingual Twitter Stylometry corpus for gender and personality profiling. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia: European Language Resources Association (ELRA). 1632-1637. ISBN: 978-2-9517408-9-1.

Contents

Chapter 1

Introduction

This technical report illustrates the creation of the TwiSty corpus (see Chapter 2) and provides detailed statistics of the corpus (see Chapter 3). The corpus was created by scraping Twitter profiles, following an idea originally proposed for English (Plank & Hovy, 2015). In the current chapter, we will first introduce author profiling and then discuss the personality type indicators used in the corpus.

1.1 Author profiling

Personality prediction based on the writing style of an author is a task belonging to the field of author profiling. Despite a growing amount of research attention (Celli *et al.* , 2014; Rangel *et al.* , 2015), computational personality recognition is hampered by the limited availability of labeled data (Nowson & Gill, 2014). Many early existing data sets contain written essays of a certain topic, which are written in highly canonical language. Such controlled settings inhibit the expression of individual traits much more than spontaneous language. As such data is hard to obtain, only limited amounts were available.

With the availability of social media text, recent efforts shifted toward using such data (Schwartz *et al.* , 2013b; Schwartz *et al.* , 2013a; Park *et al.* , 2015; Kosinski *et al.* , 2015). For example, Kosinski *et al.* (2015) collected a large amount of social media data with Big Five annotations through a tailored Facebook app. Another approach, suggested by Plank & Hovy (2015), is to use the large amounts of textual data voluntarily produced on social media (i.e., Twitter) together with self-assessed MBTI type, to collect large amounts of labeled data. As in most existing data collections, the labeling is based on the self-testing of the authors based on publicly available tests, and may contain noise if the questions of the test were not answered truthfully or if the test taken was not a good predictor of personality type.

Prior work focused almost exclusively on English, a well-represented language on Twitter. English is in fact the most frequent language on Twitter. Figure 1.1 shows the language distribution found in a sample of 65M tweets (randomly sampled over 2013). We see that 45% of all data is estimated to be English[1]. This ranking of languages is similar to what has been reported earlier (Baldwin *et al.* , 2013) and it remains rather stable if we use a larger sample.

The TwiSty corpus is created to fulfill the need for a large-scale, publicly available,

[1]We estimate the language distribution by running langid (Lui & Baldwin, 2012).

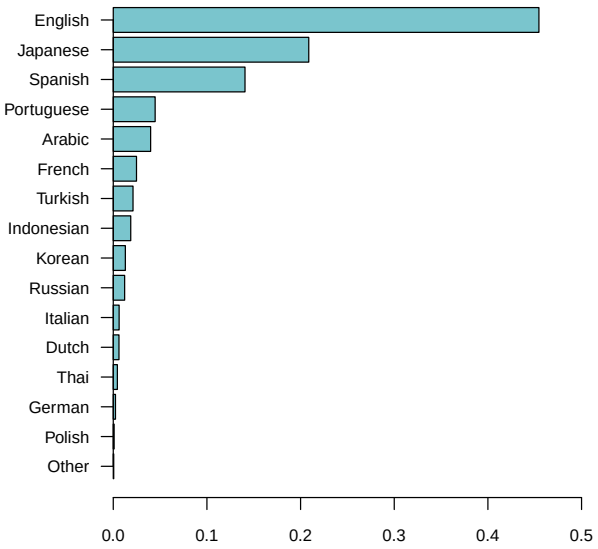


Figure 1.1: Distribution of languages (% of tweets) estimated from a 65M tweets sample (from 2013).

multilingual corpus of social media text for author profiling. It contains personality and gender annotations for a total of 18,168 authors spanning six languages (Dutch, German, French, Italian, Portuguese and Spanish).

1.2 Personality

For this corpus, we use the Myers-Briggs Type Indicator (MBTI) framework for personality (Briggs Myers & Myers, 2010). Although it is less used in psychological and engineering research (Gulati *et al.* , 2015), because of some controversy over its validity (McCrae & Costa, 1989) compared to, e.g., the Big Five (Goldberg, 1990), it is more used by the general public. This makes it particlarly well-suited to gather data for this corpus.

In the MBTI framework, one's personality type exists of four letters that each indicate a side of a dichotomy.

E/I Extraversion or Introversion

S/N Sensing or iNtuition

T/F Thinking or Feeling

J/P Judging or Perceiving

There are thus sixteen different personality types (e.g. INFJ, ESTP). By collecting personality types of Twitter users (even if self-reported on the basis of a test) together with the text of their tweets, we can investigate linguistic cues that can be linked to different personality types.

Chapter 2

Steps Taken

This chapter describes the entire process of the data collection that we implemented for each of the languages.

2.1 Twitter Mining

Because we wanted to download as many potential MBTI profiles as possible, using the Twitter API was not an option as it only allows you to search a few days back in time. Manual web search is the only way to find the entire search history, but it required us to scroll down to force loading additional tweets that matcedh our search. After loading the whole page, we could download the complete HTML source[1].

The queries we used always consisted of an MBTI type and a language-specific context word. The MBTI types form distinctive four-letter words and are thus mostly unambiguous. The context words (see Table 2.1) were chosen to be typical of each language while also being broad enough to capture as many tweets as possible. All context words were verified by a native speaker of the language.

Language	Context Words	Date Mined
German	*ich, bist, Persönlichkeit, dass*	03/11/2015
Italian	*che, fatto, sono, personalità*	21/12/2015
Dutch	*ik, jij, het, persoonlijkheid*	28/10/2015
French	*suis, c'est, personnalité*	18/11/2015
Portuguese	*sou, personalidade*	01/12/2015
Spanish	*soy, tengo, personalidad*	18/11/2015

Table 2.1: Specific context words and date of mining for each language

[1]When doing this yourself, make sure to download the complete source of the page: if you only download the HTML, you might be limited to only the tweets that were present when you first performed the search, not the extra tweets that were loaded by scrolling. We used Google Chrome as our browser of choice.

2.2 Clean-Up

2.2.1 Relevance

The combinations of the MBTI types with the context words were converted into csv files per MBTI type, where each tweet is on one line. For each tweet, the text, creation date and author details (Twitter id, handle and name) were stored. Two extra columns were provided to indicate the relevance of the tweet (i.e. a value of 0 or 1 indicating whether the tweet describes the MBTI type of its author in the appropriate language in the appropriate file) and when relevance is 1, the gender of the author (M or F). It's important to note that when more than one MBTI type is mentioned, the tweet was only marked relevant in the file of the type describing the author. If no gender could be deduced from the Twitter user's name or handle, we looked at the user's profile picture and description on Twitter. If this also did not provide us with a gender, e.g. because it was a company account, then the gender field was left open and the author was not used.

2.2.2 Gender

In a second phase of the clean-up, all relevant tweets were gathered in one csv file, ordered on author id, mentioning the MBTI type as well. When there were multiple tweets describing the same author, we checked the MBTI types of those tweets. If the type was the same for all tweets, we kept one of them and deleted the others, thus keeping one record for each author. If the type was not the same for all tweets, we reread the tweets to check if we perhaps made a mistake before or to solve possible ambiguities by looking at all tweets by this author. If there was any doubt about the type the author reported to belonged to, we discarded all tweets by this author. All authors that were not marked with a gender were deleted as well. For the bigger languages (FR, PT, ES), this final phase of the clean-up (checking double authors) was automated.

Statistics of the distribution of tweets and authors over the different MBTI personality types are given below in Table 2.3. The distributions of author gender can be found in Section 3.

Note that two personality types are frequent misspellings of common words in some languages. The most common misspelling is *infp* for *info* which occurs in all languages. Also, *estp* is a frequent misspelling of *esto* in Spanish. These misspellings explain the large discrepancy between potential and confirmed profiles in Table 2.3.

2.3 Download Tweets

We then downloaded all retrievable tweets of each user using the TwitterSearch[2] python package. There is a theoretical maximum of 3250 tweets to be downloaded, however we often found fewer tweets per author because we discard all retweets and because the author might not have written that many tweets.

Table 2.2 provides statistics on how many tweets we were able to download for each language.

[2] https://github.com/ckoepp/TwitterSearch

	Total	Mean	SD	Median	Min.	Max.	Date Mined
German	952,549	2,318	819	2,628	4	3,238	18/11/2015
Italian	932,785	1,904	912	2,146	1	3,242	01/02/2016
Dutch	2,083,484	2,083	963	2,426	3	3,215	12/11/2015
French	2,786,589	1,983	932	2,254	1	3,243	06/12/2015
Portuguese	8,833,132	2,160	878	2,456	1	3,249	19/01/2016
Spanish	18,547,622	1,722	952	1,930	1	3,244	27/01/2016

Table 2.2: Tweet counts per language before language identification.

	German		Italian		Dutch		French		Portugese		Spanish	
	B	A	B	A	B	A	B	A	B	A	B	A
ENFJ	50	18	90	20	155	91	208	75	519	229	1,027	638
ENFP	84	40	214	32	458	210	401	163	1,264	498	2,494	1,497
ENTJ	69	14	53	20	122	32	143	57	318	124	855	539
ENTP	83	26	63	19	293	101	356	99	724	224	1,187	631
ESFJ	19	8	28	8	101	47	60	24	261	123	711	465
ESFP	16	3	19	3	150	54	89	36	347	162	1,093	816
ESTJ	12	4	130	6	125	50	104	26	195	86	744	479
ESTP	17	5	137	1	106	35	470	24	201	78	2,154	313
INFJ	194	48	310	89	173	52	526	134	1,432	430	1,487	700
INFP	360	95	365	81	287	84	855	259	2,409	668	3,138	1,439
INTJ	149	38	516	89	197	58	584	160	1,798	478	2,113	900
INTP	198	60	413	71	189	68	593	171	1,904	445	1,790	751
ISFJ	40	10	27	15	57	24	104	50	373	148	468	319
ISFP	58	16	64	15	61	26	118	46	340	117	897	532
ISTJ	46	12	51	13	119	43	153	46	478	169	975	442
ISTP	62	14	49	8	98	25	218	35	351	111	598	311
Total	1,457	411	2,529	490	2,691	1,000	4,982	1,405	12,914	4,090	21,731	10,772

Table 2.3: Per category and total counts of tweets describing the author's MBTI profile. (B = before clean-up, A = after clean-up)

It is interesting to see that the corpus size for these languages mostly follows their occurrence on Twitter (cf. Figure 1.1) with only Italian and Dutch switching place in the ranking.

2.4 Language Identification

Since many Twitter users (esp. non-English users) employ more than one language (e.g. mother tongue and English), it is not sufficient to know which language a user speaks. Language identification on each tweet was performed to ensure monolinguality of our corpus. Lui & Baldwin (2014) report a majority voting approach with three language identifiers to work well for Twitter messages. We followed their approach with minor changes[3]. For that reason, the tweets were temporarily considered in a text-only form by stripping all mentions, hashtags and urls. Tweet clean-up was performed using the *get_text_cleaned* function in the *tweet_utils.py* script by Timothy Renner[4].

The three language identifiers we used are listed in Table 2.4.

Tool	Citation	Lang.	Github
langid.py	(Lui & Baldwin, 2012)	97	/saffsd/langid.py
langdetect[5]	(Nakatani, 2010)	53	/shuyo/language-detection
ldig[6]	(Nakatani, 2012)	17	/shuyo/ldig

Table 2.4: Tools used for language identification.

The results of this process can be found in Table 2.5.

	Total	Confirmed	% Confirmed
German	952,549	713,744	74.9
Italian	932,785	658,332	70.6
Dutch	2,083,484	1,541,259	74.0
French	2,786,589	1,995,865	71.6
Portuguese	8,833,132	6,353,763	71.9
Spanish	18,547,622	13,493,445	72.8

Table 2.5: Tweet counts per language before and after language identification.

2.5 Corpus structure

We created a json corpus file for each language. Each file was a dictionary with the user ids (as taken from Twitter) as keys. The value of each item was a dictionary containing the following information on the user: Twitter id, MBTI type, gender, list of tweet ids

[3] We replaced CLD2 (McCandless, 2010) with a different language identifier for reasons of availability.

[4] Available on github: https://gist.github.com/timothyrenner/dd487b9fd8081530509c.

[5] langdetect is originally written in Java. We used the python bindings from https://pypi.python.org/pypi/langdetect/1.0.1.

[6] ldig detects significantly less languages but it includes all the ones we are working with and was developed especially for Twitter.

```
{user_id1 :
        {'user_id': user_id1,
        'mbti': 'ESTP',
        'gender':'M',
        'confirmed_tweet_ids': [tweet_id1, tweet_id2, tweet_id4],
        'other_tweet_ids': [tweet_id3, tweet_id5]
        }
}
```

Figure 2.1: Corpus structure in json.

with confirmed language, and the list of all other tweet ids we considered from this user. Figure 2.1 shows what this would look like.

Chapter 3

Details per language

3.1 TwiSty-DE: German

type	male	female	total
ENFJ	4	14	18
ENFP	16	24	40
ENTJ	10	4	14
ENTP	17	9	26
ESFJ	2	6	8
ESFP	0	3	3
ESTJ	2	2	4
ESTP	2	3	5
INFJ	12	36	48
INFP	42	53	95
INTJ	16	22	38
INTP	34	26	60
ISFJ	2	8	10
ISFP	6	10	16
ISTJ	8	4	12
ISTP	13	1	14
Total	186	225	411

Table 3.1: Gender and MBTI type distribution for TwiSty-DE.

E	118	I	293
N	339	S	72
F	238	T	173
J	152	P	259

Table 3.2: MBTI trait distribution for TwiSty-DE.

3.2 TwiSty-IT: Italian

type	male	female	total
ENFJ	8	12	20
ENFP	17	15	32
ENTJ	10	10	20
ENTP	10	9	19
ESFJ	2	6	8
ESFP	1	2	3
ESTJ	2	4	6
ESTP	1	0	1
INFJ	28	61	89
INFP	23	58	81
INTJ	27	62	89
INTP	28	43	71
ISFJ	5	10	15
ISFP	4	11	15
ISTJ	5	8	13
ISTP	4	4	8
Total	175	315	490

Table 3.3: Gender and MBTI type distribution for TwiSty-IT.

E	109	I	381
N	421	S	69
F	263	T	227
J	260	P	230

Table 3.4: MBTI trait distribution for TwiSty-IT.

3.3 TwiSty-NL: Dutch

type	male	female	total
ENFJ	34	57	91
ENFP	93	117	210
ENTJ	22	10	32
ENTP	67	34	101
ESFJ	14	33	47
ESFP	16	38	54
ESTJ	22	28	50
ESTP	18	17	35
INFJ	21	31	52
INFP	42	42	84
INTJ	34	24	58
INTP	51	17	68
ISFJ	4	20	24
ISFP	8	18	26
ISTJ	23	20	43
ISTP	14	11	25
Total	483	517	1000

Table 3.5: Gender and MBTI type distribution for TwiSty-NL.

E	620	I	380
N	696	S	304
F	588	T	412
J	397	P	603

Table 3.6: MBTI trait distribution for TwiSty-NL.

3.4 TwiSty-FR: French

type	male	female	total
ENFJ	21	54	75
ENFP	67	96	163
ENTJ	33	24	57
ENTP	57	42	99
ESFJ	5	19	24
ESFP	16	20	36
ESTJ	14	12	26
ESTP	18	6	24
INFJ	42	92	134
INFP	79	180	259
INTJ	72	88	160
INTP	84	87	171
ISFJ	15	35	50
ISFP	20	26	46
ISTJ	17	29	46
ISTP	16	19	35
Total	576	829	1,405

Table 3.7: Gender and MBTI type distribution for TwiSty-FR.

E	504	I	901
N	1,118	S	287
F	787	T	618
J	572	P	833

Table 3.8: MBTI trait distribution for TwiSty-FR.

3.5 TwiSty-PT: Portuguese

type	male	female	total
ENFJ	97	132	229
ENFP	197	301	498
ENTJ	67	57	124
ENTP	120	104	224
ESFJ	38	85	123
ESFP	49	113	162
ESTJ	40	46	86
ESTP	41	37	78
INFJ	139	291	430
INFP	245	423	668
INTJ	194	284	478
INTP	187	258	445
ISFJ	46	102	148
ISFP	40	77	117
ISTJ	72	97	169
ISTP	50	61	111
Total	1,622	2,468	4,090

Table 3.9: Gender and MBTI type distribution for TwiSty-PT.

E	1,524	I	2,566
N	3,096	S	994
F	2,375	T	1,715
J	1,787	P	2,303

Table 3.10: MBTI trait distribution for TwiSty-PT.

3.6 TwiSty-ES: Spanish

type	male	female	total
ENFJ	240	398	638
ENFP	518	979	1,497
ENTJ	312	227	539
ENTP	352	279	631
ESFJ	156	309	465
ESFP	273	543	816
ESTJ	261	218	479
ESTP	184	129	313
INFJ	239	461	700
INFP	499	940	1,439
INTJ	467	433	900
INTP	365	386	751
ISFJ	106	213	319
ISFP	198	334	532
ISTJ	262	180	442
ISTP	181	130	311
Total	4,613	6,159	10,772

Table 3.11: Gender and MBTI type distribution for TwiSty-ES.

E	5,378	I	5,394
N	7,095	S	3,677
F	6,406	T	4,366
J	4,482	P	6,290

Table 3.12: MBTI trait distribution for TwiSty-ES.

Acknowledgements

This research is supported by a doctoral grant from the FWO Research Council - Flanders for the first author. We thank Guy De Pauw and Tom De Smedt for technical support. Part of this research was carried out in the framework of the AMiCA (IWT SBO-project 120007) project, funded by the Flemish government agency for Innovation by Science and Technology (IWT).

References

Baldwin, Timothy, Cook, Paul, Lui, Marco, MacKinlay, Andrew, & Wang, Li. 2013. How Noisy Social Media Text, How Diffrnt Social Media Sources? *Pages 356–364 of: Proceedings of the International Joint Conference on Natural Language Processing*.

Briggs Myers, Isabel, & Myers, Peter. 2010. *Gifts differing: Understanding personality type*. Nicholas Brealey Publishing.

Celli, Fabio, Lepri, Bruno, Biel, Joan-Isaac, Gatica-Perez, Daniel, Riccardi, Giuseppe, & Pianesi, Fabio. 2014. The workshop on computational personality recognition 2014. *Pages 1245–1246 of: Proceedings of the ACM International Conference on Multimedia*. ACM, Orlando, FL, USA.

Goldberg, Lewis R. 1990. An Alternative "Description of Personality": the Big-Five factor structure. *Journal of personality and social psychology*, **59**(6), 1216.

Gulati, Jayati, Bhardwaj, Priya, & Suri, Bharti. 2015. Comparative study of personality models in software engineering. *Pages 209–216 of: Proceedings of the Third International Symposium on Women in Computing and Informatics (WCI'15)*. ACM.

Kosinski, M., Matz, S., Gosling, S., Popov, V., & Stillwell, D. 2015. Facebook as a Social Science Research Tool: Opportunities, Challenges, Ethical Considerations and Practical Guidelines. *American Psychologist*, **70**(6), 543–556.

Lui, Marco, & Baldwin, Timothy. 2012. langid.py: An off-the-shelf language identification tool. *Pages 25–30 of: Proceedings of the ACL 2012 system demonstrations*. Jeju, Korea: ACL.

Lui, Marco, & Baldwin, Timothy. 2014. Accurate language identification of Twitter messages. *Pages 17–25 of: Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*. Gothenburg, Sweden: ACL.

McCandless, Michael. 2010. *Accuracy and performance of Google's compact language detector*. http://blog.mikemccandless.com/2011/10/accuracy-performance-of-googles.html.

McCrae, R., & Costa, P. 1989. Reinterpreting the Myers-Briggs Type Indicators from the perspective of the five-factor model of personality. *Journal of Personality*, **57**, 17–40.

Nakatani, Shuyo. 2010. *Language Detection Library for Java*. https://github.com/shuyo/language-detection.

Nakatani, Shuyo. 2012. *Short text language detection with infinity-gram*. https://shuyo.wordpress.com/2012/05/17/short-text-language-detection-with-infinity-gram/.

Nowson, Scott, & Gill, Alastair J. 2014. Look! Who's Talking?: Projection of Extraversion Across Different Social Contexts. *Pages 23–26 of: Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition*. ACM.

Park, Greg, Schwartz, H Andrew, Eichstaedt, Johannes C, Kern, Margaret L, Stillwell, David J, Kosinski, Michal, Ungar, Lyle H, & Seligman, Martin EP. 2015. Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, **108**(6).

Plank, Barbara, & Hovy, Dirk. 2015. Personality Traits on Twitter–or–How to Get 1,500 Personality Tests in a Week. *In: Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*. Lisbon, Portugal: ACL.

Rangel, Francisco, Celli, Fabio, Rosso, Paolo, Potthast, Martin, Stein, Benno, & Daelemans, Walter. 2015. Overview of the 3rd Author Profiling Task at PAN 2015. *In: CLEF 2015 Working Notes*. Toulouse, France: CEUR.

Schwartz, Hansen Andrew, Eichstaedt, Johannes C, Kern, Margaret L, Dziurzynski, Lukasz, Ramones, Stephanie M, Agrawal, Megha, Shah, Achal, Kosinski, Michal, Stillwell, David, Seligman, Martin EP, *et al.* . 2013a. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, **8**(9).

Schwartz, Hansen Andrew, Eichstaedt, Johannes C, Dziurzynski, Lukasz, Kern, Margaret L, Blanco, Eduardo, Kosinski, Michal, Stillwell, David, Seligman, Martin EP, & Ungar, Lyle H. 2013b. Toward Personality Insights from Language Exploration in Social Media. *In: AAAI Spring Symposium: Analyzing Microtext*. AAAI.

Verhoeven, Ben, Daelemans, Walter, & Plank, Barbara. 2016. TwiSty: A multilingual Twitter Stylometry corpus for personality and gender profiling. *In: Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia: European Language Resources Association (ELRA).

List of Tables

List of Figures