GeoCorpora: Building a Corpus to Test and Train Microblog Geoparsers
–Supplementary Online Materials–

Jan Oliver Wallgrün[a*], Morteza Karimzadeh[b*], Alan M. MacEachren[b*], and Scott Pezanowski[b*]

[a]*Independent Researcher (Affiliated with GeoVISTA Center and ChoroPhronesis, Pennsylvania State University), Ahrensburg, Germany*; [b]*Department of Geography, GeoVISTA Center, Pennsylvania State University, University Park, PA, USA*

Appendix A. Instructions to AMT Workers

In the instruction part of the web page that workers were presented with, we show an animated image to illustrate the highlighting process together with the following instructions:

(1) Please make sure that words that make up a single place name are highlighted together. For example, "New York" in the example above should appear as New York when you are done, NOT as New York.
(2) If two related place names occur separated by a comma as in "Boston, MA", treat these as individual place names and mark them separately (see example above).
(3) Make sure you do not include spaces or punctuation marks before or after a place name in your selection unless you are sure the punctuation is part of the place name, as it might be in Hawaiian (e.g., 'Aiea, Hawai'i).

We also provided additional instructions (with examples) on what should be marked and what should not. We here list these instructions and examples exactly as they appeared on the web page presented to the AMT workers.

Things that should be marked are:

(1) Any named town, city, county, state, or country (e.g., Los Angeles, Jefferson County, NY, Italy)
(2) Named buildings (e.g., Eiffel Tower, Dodgers Stadium, Alcatraz, James J. Ferris High School)
(3) Named areas (e.g., Grand Canyon, Pacific Ocean, Washington Mall, Hyde Park, Mount Washington)
(4) Street and highway names (Atherton Street, 1st Ave, Highway 1, I-70)

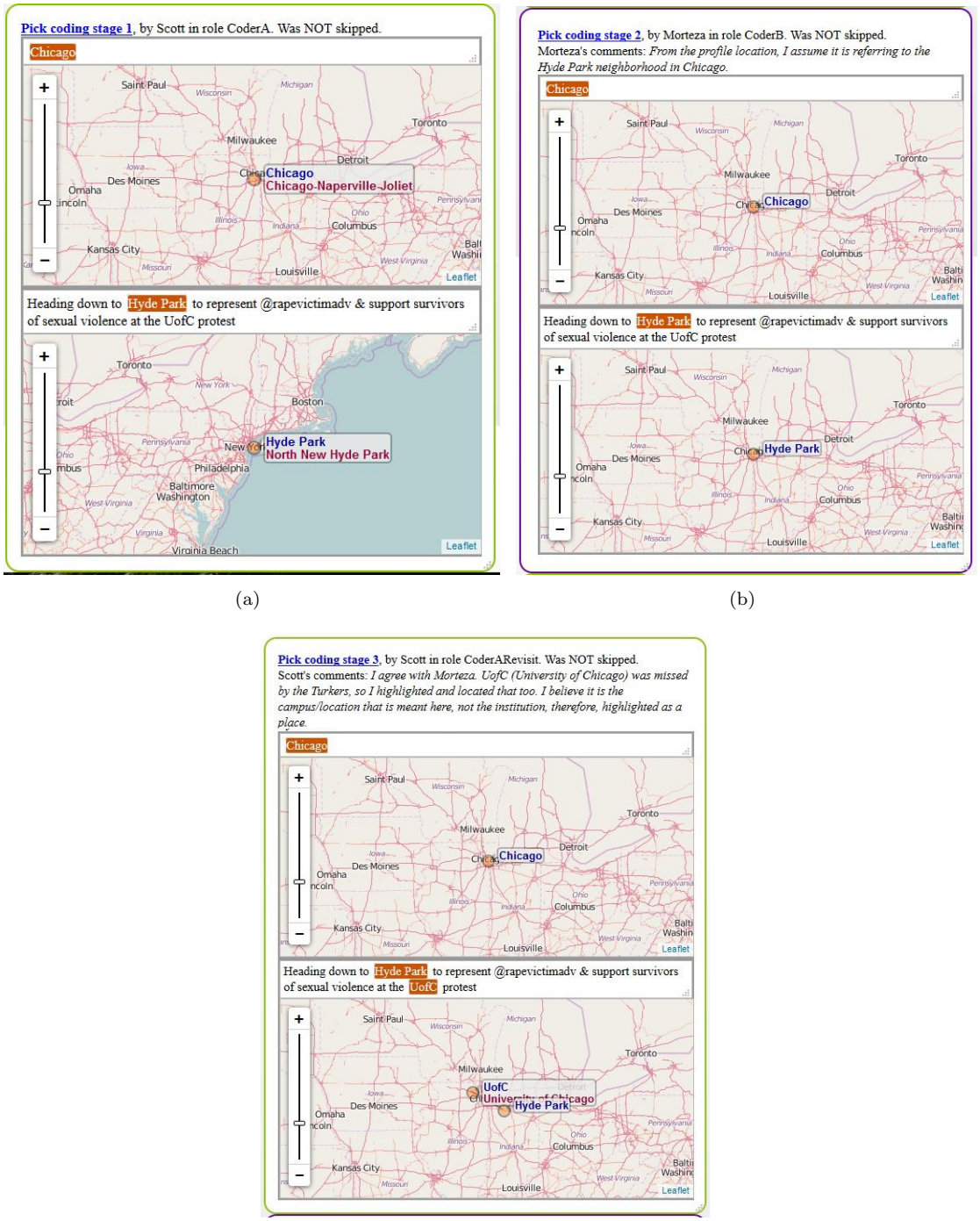
*Corresponding authors. Email: {wallgrun,karimzadeh,maceachren,scottpez}@psu.edu

(5) Abbreviations and nicknames of place names (e.g. MA for Massachusetts, Big Apple for New York City, or Chi-Town for Chicago)
(6) Place names inside hashtags (e.g. #prayforindonesia); please mark only the place name(s) contained in the hashtag, so just Indonesia in this example

Things not to be marked are:

(1) Businesses (Starbucks, Microsoft, Texas Steak House, Baltimore Ravens), even if their name contains a place name
(2) Organizations (Lutheran Church, British Red Cross, United Nations, Grand Canyon Historical Society), even if their name contains a place name
(3) Place names used as descriptors / adjectives (U.S. dollar, Philadelphia cheesesteak, South Sudan's rebels)
(4) Vague qualifiers (e.g. south, upper, central) of place names (upper Manhattan, southern Pennsylvania, central Europe) unless they are part of the proper place name as in "South Africa"; the proper place name itself (e.g. Manhattan in "upper Manhattan") should be marked though
(5) References to locations without a proper place name; e.g. in the following, include "Hudson River" but do NOT include "south shore":"Flash flood watch for the Hudson River has now been trimmed back to just include the south shore through 9am."
(6) Kind of place when not part of the proper name (e.g. "Island" in "Long Island" is part of the place name and should be marked; but "island" in "island of Sabang" is not part of the name, so only the place name itself (Sabang) should be marked)

Appendix B. Annotation History Window of Geo-Annotator



(a) (b)

(c)

Figure B1. Annotation history window where annotators can see each others annotations for a tweet.

Appendix C. Corpus Statistics

C.1. *Geographic Distribution of Identified Place Names*

Table C1 summarizes the geographical distribution of the place names identified in the corpus (not counting problematic cases) on a per-continent basis[1].

Table C1. Geographical distribution of identified place names in corpus

Continent	% of unproblematic place names by continent
Africa	11.59%
Asia	25.30%
Europe	13.66%
North America	46.21%
Oceania	2.16%
South America	1.08%

Table C2 lists the 15 countries containing the highest numbers of unproblematic place names in the corpus.

Table C2. 15 countries with highest place name numbers in corpus

Country	% of unproblematic place names by country
United States	40.83%
United Kingdom	5.69%
Syria	3.09%
India	2.80%
Ukraine	2.68%
Israel	1.98%
Iran	1.90%
Japan	1.90%
Nigeria	1.73%
Sierra Leone	1.65%
Pakistan	1.57%
Palestine	1.49%
Malaysia	1.40%
Iraq	1.40%
China	1.24%

[1]105 place names in the corpus were not captured by this statistic because GeoNames did not provide continent information for these.

C.2. *Distribution of Entity Types of Identified Place Names*

Table C3 lists the 20 most frequent GeoNames feature types[2] over the identified unproblematic place names in the corpus. Not all toponyms in GeoNames have a feature code assigned though, and these had to be excluded from the statistics.

Table C3. Distribution of unproblematic place names by GeoNames feature type

Feature type	% of unproblematic place names by feature type
Independent political entity (typically country names)	38.72%
First-order administrative division (typically state or province names)	13.32%
Populated place (cities, towns)	11.51%
Seat of second-order administrative division (county or governorate capital)	8.37%
Seat of first-order administrative division (state or province capital)	5.44%
Capital of a political entity	4.99%
Second-order administrative division (county or governorate)	2.97%
Continent	2.97%
Region (defined or loosely defined named area)	0.95%
Building	0.83%
Semi-independent political entity (e.g. Palestine)	0.62%
Stream	0.58%
Amusement park	0.58%
Economic region	0.50%
Seat of third-order administrative region (seat of a third-order admin division such as a borough)	0.50%
Section of populated place (Urban neighborhoods such as "Hyde Park" in Chicago)	0.41%
Airport	0.41%
Island	0.41%
School	0.37%
Park	0.33%

[2]http://www.geonames.org/export/codes.html

Appendix D. Examples of Tweets with AMT Disagreement

Table D1 lists examples of tweets that contained place reference (candidates) that led to significant (> 30%) AMT worker disagreement. For each case it lists the disagreement category (discussed in Section 3.2 of the article) and provides a brief explanation in the Comment column.

Table D1. Examples of tweets with AMT disagreement

Tweet text	Disagreement category	Comment
"#furgeson I didn't know it was, "wear a hoodie to the riot" day. Damn pullover!" (Tweet ID 537064161977434113)	TweetDef	Misspelling of Ferguson caused several workers to not recognize it as a place reference.
"Ik I'm from vt and we don't get tornados.. Buuut this is the most weak tornado I've ever seen" (Tweet ID 509824099716444160)	TweetDef	The fact that "vt" is spelled in all lower-case letters most likely resulted in part of the workers overlooking the abbreviation for the US state of Vermont (VT).
"Clashes in Syria near Golan Heights - BEIRUT (AP) Syrian troops battled rebels near the Israeli-occupied Golan H... http://t.co/F1JoouZgag" (Tweet ID 530707574472343552)	TweetDef	Tweet contains a second reference to the Golan Heights but it has been truncated to "Golan H" resulting in some workers marking just "Golan" or nothing at all.
"RT @ABC7: Protest unrest! Police fire #teargas in #Ferguson, MO days after the shooting of #MichaelBrown. At 11p @abc7 https://t.co/GH2ZH7f" (Tweet ID 499812800324063233)	GroupingErr	In this example, we have a typical city-state pair (Ferguson and MO for Missouri) separated by a comma, so both city and state should have been marked separately. However, some workers highlighted the entire phrase together.
"RT @paulturkey: 2,000 people turned out in the wind and rain today in Arklow Co Wicklow to protest against the water charges well done all" (Tweet ID 529227404452327424)	GroupingErr	Unfamiliarity of many workers with the involved places, a tweet deficiency in the form of a missing comma clearly separating the two involved entities, the use of the abbreviation Co for Country, and potentially unfamiliarity with the British and Irish way of putting "County" in front of a county name rather than behind it (as, for instance, in the U.S.) have strongly contributed to the varying AMT results we got for this tweet.

Tweet text	Disagreement category	Comment
"Hickory Ridge Mall wasn't da same after the tornado ????" (Tweet ID 536611528662061056)	MissedPlaceErr	"Hickory Ridge Mall" was not marked by several workers. Since this entity is the subject of the sentence and begins the sentence, our interpretation here is that they did not take into account our rule that named buildings should be included, or they did not consider a mall to be a "building" since it is not a prototype of that category due to its size.
"RT @EuromaidanPR: Comparing the uniforms of "rebels" in the #Donetsk region and of #Russian SOF in #Grozny @Fbeyeee —EMPR http://t.co/8NJpM" (Tweet ID 540582148907220992)	MissedPlaceErr	Several workers did not mark the mention of the city of "Grozny" in the tweet, presumably because of not being familiar with the name. Another explanation here could be that these workers did not take into account the rule to include place names that are part of a hashtag, but at least some of them marked "Donetsk" which also is part of a hashtag, so unfamiliarity seems like the most likely explanation.
"RT @VacciNewsNet: Vaccination prevents an estimated 2 million cases of the flu every year in the US #GetAFluShot http://t.co/9s611ys3rW" (Tweet ID 512468244515786752)	MissedPlaceErr	Surprisingly, more than a third of the workers tagging this tweet did not mark the occurrence of "US". We believe that the two-letter abbreviation in all upper-case letters without periods, almost at the end of the tweet, and only followed by a hashtag and an URL made these workers simply overlook the place name.
"#News: At least 1,119 Iraqis died in violence in September, toll excludes slayings by IS group: The U.N. http://t.co/QwphtAUm2B #TU" (Tweet ID 517254978889457665)	NotAPlaceErr	While we did not provide an explicit rule for this case, our view is that this is a reference to a particular group of people rather than a direct reference to a place and, hence, should not be tagged.
"The number one viral killer in Mississippi and the U.S. isn't Ebola, or enterovirus, or West Nile. It's flu. #flushot http://t.co/a5OLbh60KT" (Tweet ID 514777489651294208)	NotAPlaceErr	West Nile here clearly is first of all a reference to the disease and not a direct reference to a place. Therefore, it should not have been marked.
"RT @PATHdrugdev: Waiting on the #turkey? Catch up on #globalhealth news w updates on #HIV #malaria #pneumonia & more: http://t.co/UYCuJvIxE" (Tweet ID 537999680958124032)	NotAPlaceErr	Roughly half of the AMT workers marked "turkey" as a place name. While going by the text itself, it is somewhat possible interpretation that "turkey" refers to the country, the fact that the tweet was written on Thanksgiving day (for the U.S.) makes it clear that it is a reference to the bird instead and shouldnt have been marked.

Tweet text	Disagreement category	Comment
"It was a mini riot at toledos mall for the Columbia 11's ticket" (Tweet ID 545385171474993152)	NotAPlaceErr	This is somewhat tricky example but almost 70% of the workers tagging this tweet got it right that "Columbia 11" is the name of shoe and, hence, "Columbia" should not be marked here.
"RT @DFID_UK: Summary of pledges from countries, charities & companies made at the Defeating Ebola conference now online: http://t.co/VWTTKd" (Tweet ID 517737521196064768)	TweetUserID	After discussing these cases, we decided that twitter IDs, first of all, are references to persons or organizations and not direct references to places, even if they include place names. Therefore, we established the additional rule that place references in twitter IDs should never be tagged and considered as place mentions.
"BACK TO THE FUTURE ARTIST SHOWCASE — AUG. 24th @ FIRE & ICE! 312 MARKET ST. 5pm - 10pm! Tickets 10&15 at the door! ?? BE HERE!!" (Tweet ID 496371470722150400)	AddressRelated	Some AMT workers did not mark "312 MARKET ST." at all, others marked only "MARKET ST.", or others marked the entire phrase. It is our view that an address constitutes a named place and, hence, should be marked and that the house number and street name should be considered as a single place reference to be marked together.
"RT @Watcherone: South Sudan rebels have killed several Uganda soldiers in the Upper Nile in renewed fighting in the country." (Tweet ID 521571268362252288)	DescAdjUsage	Both "South Sudan" and "Uganda" are noun adjuncts used as descriptors for groups of people in this tweet. Such noun adjuncts and other forms of adjectival use can refer to the place, the people, the language, or to the government of a place. In the example here, the purpose is most likely is to describe the origin of the respective group of people but it could also describe their current location.
"West Bank violence erupts as Gaza conflict rages: The Israeli military shot and killed two Palestinian men and... http://t.co/s7GFn1hOfw" (Tweet ID 492640920748044289)	DescAdjUsage	In the phrase "West Bank violence", "West Bank" seems to convey the location where the violence (an ongoing event) is taking place, so could be seen as a place reference that should be tagged. "Gaza" in "Gaza conflict" is an established name for an ongoing event, but one that is not entirely restricted to the area given by the place name.

Tweet text	Disagreement category	Comment
"RT @miamivice_22: It is a photo at the time of the Great Hanshin-Awaji Earthquake. Picture hell. http://t.co/rwzzhexLgn" (Tweet ID 556164086677774338)	DescAdjUsage	This is another example of a place name linked to an event. "Hanshin-Awaji" can be seen as describing the location of this particular earthquake as in "the great earthquake in Hanshin-Awaji". Historic natural disasters limited to a certain area often can be so closely associated with the area struck, that they can also be considered a reference to a place as much as a reference to the event.
"RT @RealJamesWoods: Rebels capture Yemen presidential palace http://t.co/55DnT5JBBA The Obama foreign policy failures are simply staggering" (Tweet ID 557629977110523906)	DescAdjUsage	Specific location or geographic entity is referred to with a qualifier that contains a proper name (potentially a proper place name). We got mixed results from the AMT workers in these examples ranging from not marking anything, marking just the proper place name, to marking the entire expressions including the qualifier.
"RT @UberFacts: In June 2008, a tornado hit the Kansas State University campus destroying only one building - Their Wind Erosion Lab." (Tweet ID 510956886733385728)	DescAdjUsage	Specific location or geographic entity is referred to with a qualifier that contains a proper name (potentially a proper place name). We got mixed results from the AMT workers in these examples ranging from not marking anything, marking just the proper place name, to marking the entire expressions including the qualifier.
"Here's a look at the traffic backup following a truck fire on the Mississippi River Bridge this morning, eastbound. http://t.co/9ri5KrJEYW" (Tweet ID 487627602744463363)	DescAdjUsage	This example shows that it can also be hard to say when a proper name is a qualifier and when it is part of a the proper name itself as "Mississippi River" in this example.
"@OleMissMBB: Ole Miss knocks off Mississippi State 79-73. Summers with 22 points for the Rebels, all in the second half. @savannahrae07" (Tweet ID 560652029417496576)	Organization-Related	"Mississippi State" (as well as "Ole Miss") refers to a sports team that is associated with a larger organization (Mississippi State University) which in turn may be loosely associated with a particular geographic area (e.g. the universitys main campus). As such, it is not a direct place reference and should not be tagged.
"Iran: protest rally in front of Gilan governor office against the shutdown of Looshan Cemnet Factory http://t.co/8wkH2239CH" (Tweet ID 549943792636133376)	Organization-Related	"Looshann Cemnet Factory" is the name of a business (misspelled) but is more closely associated with a particular place than "Mississipi State" in the previous example.

Tweet text	Disagreement category	Comment
"RT @we_support_PTI: All Pakistani's In USA - COME OUT to Protest #GoNawazGo, as the #FakePrimeMinister visits United Nations. #ImranKhan ht" (Tweet ID 515036649768890368)	Organization-Related	"United Nations" can refer to the United Nations organization or subdivision of this organization but also be seen as a reference to the United Nations Building. "United Nations" as the object to "visits" hints at this interpretation.
"Pakistan Awami Tehreek (PAT) Protest at the United Nations Secretariat http://t.co/8dUtxmkSRK @locke @AlanFisher @AJELive @azadessa" (Tweet ID 515638387001536512)	Organization-Related	"United Nations Secretariat" here is similar to "United Nations" in the previous example but the interpretation of a place is even more likely here because of using the name in connection with the spatial preposition "at".
"RT @specterm: Will smbdy have to die before ppl get the message? RT @APHL Measles outbreaks hit 18-year high in Washington state http://t." (Tweet ID 487688329643565056)	KindOf-PlaceQual	"state" is not part of the of the proper place name but it is most likely used to to distinguish it from Washington D.C., often also only referred to as Washington. Some AMT workers marked just "Washington" and some marked "Washington state" in this case.
"RT @BaltimoreParade: On May 27, 1798, at the Battle of Oulart Hill, County Wexford, 1,000 rebels under Father John Murphy annihilated... ht" (Tweet ID 521842978218721281)	KindOf-PlaceQual	This tweet contains two examples of kind of place qualifier: "Hill" in the case of "Oulart Hill" and "County" in the case of "County Wexford". The official name of the country is "County Wexford" which distinguishes it from its county town just called "Wexford". While AMT results for "County Wexford" where split half-and-half between marking just "Wexford" and marking "County Wexford", there was sufficient agreement (according to our 70% criterion) that "Oulart Hill" should be marked together.
"The Ebola crisis in west Africa is outstripping the ability of aid organisations to stem the epidemic http://t.co/kfjdT2G9cE" (Tweet ID 500417994292744193)	Vaguely-Qualified	This is an example where a cardinal direction ("west") is used as a vague qualified.
"7.3-Magnitude Earthquake in Eastern Indonesia, Tsunami Warning Issued: A 7.3-magnitude earthquake rocked the Maluku Islands in easter" (Tweet ID 533470176847486976)	Vaguely-Qualified	Another example of a cardinal direction ("Eastern") used as a vague qualified.
"RT @EuromaidanPR: Comparing the uniforms of "rebels" in the #Donetsk region and of #Russian SOF in #Grozny @Fbeyeee —EMPR http://t.co/8NJpM." (Tweet ID 540582148907220992)	Vaguely-Qualified	In this example the vagueness is introduced by the trailing "region".

Tweet text	Disagreement category	Comment
"#BREAKING M6.0 earthquake jolted the sea area near S. Sumatra Wed., the quake hit at a depth of 10km.(CENC) http://t.co/E0NkG1mbkV" (Tweet ID 545111381884694529)	Vaguely-Qualified	"S. Sumatra" could vaguely refer to the southern part of the island of Sumatra or the province called South Sumatra whose boundaries are geographically defined precisely and which is listed as a first-order administrative division in GeoNames.

Appendix E. Other Examples

Table E1. Other interesting examples from the corpus

Tweet text	Comment
"NWS in Paducah has issued a flood warning for the Mississippi River at Cape Girardeau, affecting Alexander, Jackson & Union counties in IL." (Tweet ID 511546368813322243)	Tweet with the highest number of place names (7) in a single tweet