# Building a Dataset for Possessions Identification in Text

**Carmen Banea, Xi Chen, Rada Mihalcea**
Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, Michigan
carmennb@umich.edu, chenxicx@umich.edu, mihalcea@umich.edu

**Abstract**

Just as industrialization matured from mass production to customization and personalization, so has the Web migrated from generic content to public disclosures of one's most intimately held thoughts, opinions and beliefs. This relatively new type of data is able to represent finer and more narrowly defined demographic slices. If until now researchers have primarily focused on leveraging personalized content to identify latent information such as gender, nationality, location, or age of the author, this study seeks to establish a structured way of extracting possessions, or items that people own or are entitled to, as a way to ultimately provide insights into people's behaviors and characteristics. In order to promote more research in this area, we are releasing a set of 798 possessions extracted from blog genre, where possessions are marked at different confidence levels, as well as a detailed set of guidelines to help in future annotation studies.

**Keywords:** possession identification, latent attribute extraction, author profiling

## 1. Introduction

> "Watch your thoughts, they become your words.
> Watch your words, they become your actions.
> Watch your actions, they become your habits.
> Watch your habits, they become your character.
> Watch your character, it becomes your destiny."
> *Author Unknown*

With the introduction and adoption of Web 2.0, the Internet has become a forum where users voice their opinions and feelings through comments, reviews, blogs, microblogs, status updates, and other forms of online participation. This growing and diverse unstructured stream of information blends for the first time consumer demographics, with lifestyle information and choices, user opinions and feelings, as well as mentions of items that are owned or liked by the user.

Our research is motivated by the affective-cognitive consistency model (Rosenberg, 1956; Rosenberg, 1968), a branch of cognitive consistency theory that not only hypothesizes that people are motivated to seek a coherent state both internally (at the level of thoughts, beliefs, feelings, and values) and externally (through attitudes and behaviors), but also that individuals gain more motivation in achieving a consistent state so that others perceive them to be consistent. This particular model implies that in a public setting, where others are reading and scrutinizing a person's online content, the blended data forms not only a raw user profile, but it encodes a coherent user signature, which, if mined successfully, can carve out very narrowly defined clusters of individuals who live, think and act alike.

Motivated by the potential of analyzing and evaluating people's behaviors and characteristics as they relate to their ownership of a given object, this work seeks to establish the framework for *identifying such objects*. To this end, we propose the concept of "possessions," or textual representation of items that somebody owns or is entitled to, which we define in detail in Section 2.1.

To date, research focusing on extracting latent attributes from microblogs has mostly centered around Twitter, as it is a service with a high adoption rate, where many of the users share their tweets publicly. Some of the attributes targeted for extraction are demographics related, such as gender and age (Burger and Henderson, 2006; Mukherjee and Liu, 2010; Rao et al., 2010; Burger et al., 2011; Van Durme, 2012), political affiliation (Conover et al., 2011; Cohen and Ruths, 2013), and even lifestyle choices, such as coffee preference (Pennacchiotti and Popescu, 2011). To our knowledge, this is the first study that seeks to identify object ownership.

## 2. Possessions

We start by defining possessions, and then briefly review the main considerations we established in order to produce consistent annotations.

### 2.1. What Are Possessions?

We define possessions as textual representations of physical, concrete objects that could be considered to be someone's property such as electronics, clothes, furniture, etc., or of items to which somebody is entitled to due to his / her position or social standing, such as an employee to his cubicle, or a king to his throne. Possessions, however, cannot be human beings, as people can exercise free will: "my mother" appearing in a given context does not render the denoted person a possession, despite the preceeding possessive article.

### 2.2. Considerations

Several factors need to be considered when one thinks of possessions and their attributes.

**Ownership.** We identify possessions with respect to the author of the utterance. For example "I left my laptop in the car," suggests that the writer owns a laptop; however, it is important to note that the same context sheds no light on whether the car is implicitly his / hers as well; as such, considering the limited information, the automobile is

not considered a possession. Another aspect to examine is the fact that a possession can exhibit joint ownership. For example, in the sentence "My husband and I own a beautiful house," the house is an object to which both parties are entitled, thus the object is a possession of the speaker, as well.

**Time frame.** A given object needs to be possessed by the writer at the time of the utterance in order to be considered a possession. Items owned in the past or whose current status is unknown are not considered possessions. This time frame consideration also allows accounting for negations, irrealis and sarcastic statements in text. For example statements such as "I never had a car" (negation), "I always wished I had a car" (irrealis), "Of course, I have a personal chopper!" (sarcasm), can be accurately processed as containing no possessions.

**Identifiability.** When annotating possessions, one of the main aspects to consider is the identifiability of the expressions being annotated. Let us consider the following sentence: "I left my green shoulder bag in the car." The identified item cannot be simply "bag," as that would be too ambiguous; are we talking about a backpack, a shopping bag, a beauty bag, purse, or luggage? For this reason, the shortest span we can annotate, that will also provide a precise idea of the actual item being possessed, is "shoulder bag." In addition, generic words such as "things," "items," "collections" should not be annotated, as they are not identifiable.

**Document level consistency.** Furthermore, the identification of possessions is considered with respect to the entire document. Items whose ownership status may have been unknown in the initial passages of a document, may be attributable to the writer once all the context has been taken into consideration. As such, all mentions of those items in the document are properly resolved upon a second pass.

**Concrete nouns.** Possessions are concrete nouns representing objects that can occupy a physical space. Another way to think of such objects is by considering their picturability potential. Nouns such as cup, fork, desk, computer, are concrete, and therefore potential possessions, while nouns such as love, happiness, goals, etc. are not. Another consideration is that even if an item exists only in a virtual world, such as an email, blog, document or photograph, the fact that such items are printable and therefore can become tangible, renders them potential possessions.

**Resolution scope.** While items are decomposable in terms of the constituent parts, we consider possessions with respect to the whole. A cellphone may contain a screen, microphone, case, etc., but the possession being identified is the cellphone. By the same token, body parts are not annotated because they resolve to a person, and as mentioned earlier, persons are not items in order to be considered property.

For the complete annotation guidelines, we invite the reader to consult the "Possession identification guidelines" released with this article[1].

## 3. Annotation Format

All possessions that meet the considerations touched upon in the previous section are marked in the text using an XML-based formatting schema. Each possession is identified by an $< object >$ tag, which can take several attributes:

- *value*: an expression describing the item type as found explicitly or implied from the text. Partial textual references to an item are resolved and cross-referenced to an identifiable object (as required by the identifiability constraint). All items having the same *id* also have the same *value*.
- *id*: a unique number identifying the possession within the document. It starts at 0, and is incremented every time a new possession is identified within the same document. Multiple mentions of the same possession within a document are resolved to the same id.
- *type*: "perm" (permanent) / "temp" (temporary); refers to how persistent the possession is. If a possession lasts less than one day, we consider it "temporary," otherwise, it is perceived to be "permanent." For example, if the possession is a perishable item (ice-cream, coffee) or an item that is not expected to last (ice), the type is set to "temporary."

Let us revisit the example:

> "*I left my green shoulder bag in the car.*"

Once annotated, the sentence becomes:

> I left my <object value = "shoulder bag" id = "0" type ="perm"> green shoulder bag </object> in the car.

The *span* of the annotation is "green shoulder bag," as it unequivocally establishes the object that belongs to the writer. However, the *value* of the object is devoid of personalized information, thus allowing cross owner profile analysis (both A and B may possess a shoulder bag, but only A's bag is green). The *id* of the item is 0, assuming that it is the first possession encountered in the blog, and its *type* is permanent, as it is an object that will last the owner an extended period of time.

As the data is initially annotated by two judges, and then reconciled by a third judge, at the end of the reconciliation stage two additional attributes are introduced:

- *status*: denotes the fact that the third annotator casts a confidence vote for a particular marked span by adding a $status =$ "*final*" attribute.
- *agreement*: lists the first name initial of the particular annotators who identified an object. For $agreement =$"*em*", two judges concur in their determination.

[1]Also available at `http://lit.eecs.umich.edu/research/downloads/`

## 4. Dataset

We are releasing a dataset containing annotations of 798 possessions that appeared in 27 blogs collected from the Internet between May and July 2015, from categories such as lifestyle, travel, health, weddings, shopping (see Table 1 for a more detailed breakdown). Mapping to these categories was achieved via Walmart's API[2]. Unsurprisingly, due to the personal nature of the blogs, most possessions are associated with the lifestyle category.

Several University of Michigan senior students majoring in linguistics expressed interest in participating in the annotation study. To select the right candidates, we circulated a set of annotation guidelines and two sample blogs that the candidates annotated, and we compared their answers with an in-house developed gold standard. Ultimately, based on the quality of the annotations, three candidates were selected.

The set of blogs was split into three parts, and through rotation, each subset of 9 blogs was assigned to a different annotator to reconcile, upon receiving individual annotations encoded by the other two judges. The third annotator was the only one who could specify the *status* and the *agreement* attributes in the $< object >$ tag, as mentioned in Section 3.. Out of a total of 798 annotations, 614 (76.94%) are marked with the *status="final"* attribute, indicating that the task is well defined, thus eliciting a high inter-annotator agreement.

Based on the reconciled annotations, we are releasing several versions of the dataset with different confidence levels: the gold-standard, where all three annotators agreed on the possession being marked (345 possessions / 43.23%), the silver-standard, where at least two annotators agreed (583 possessions / 73.06%), and the bronze-standard, consisting of all the annotations made (798 possessions / 100%)[3].

Overall, we encountered 798 possessions (average of 29.56 per blog); these are mostly expressed through nouns, at an average of 1.07 nouns per span. In some instances, such as "I have to brush my hair," the verb "to brush" is annotated to imply the possession of a hairbrush, despite the fact that the physical object is not directly mentioned in the sentence. 89 instances where the possession span included a verb were identified, representing 11.15% of the total number of possessions.

Among the most often encountered possessions in the gold-standard we have: house (27), prosthetic leg (16), phone (16), wedding dress (15), and car (11). For the silver-standard, additional high-frequency possessions are items such as photo (18), blog (16), and glucose monitor (10), while for the bronze-standard we have loaner car (10), gym (10) and picture (10).

Out of 12,766 nouns that occur in the blogs (identified using the automatic part-of-speech tagger included with the NLTK package[4]), only 709 of them were marked as possessions by the annotators, representing a proportion of 5.55%. This implies that despite the personal nature of blog genre, relatively very few nouns actually represent author possessions, further indicating that the task of automatically extracting possessions from text may be quite challenging.

| Category | # of blogs | # objects | Average |
|---|---|---|---|
| Lifestyle | 8 | 265 | 33.12 |
| Travel | 4 | 104 | 26.00 |
| Other | 3 | 52 | 17.33 |
| Health | 3 | 105 | 35.00 |
| Wedding | 2 | 69 | 34.50 |
| DIY | 1 | 32 | 32.00 |
| Real Estate | 1 | 6 | 6.00 |
| Parenting | 1 | 14 | 14.00 |
| Pets | 1 | 23 | 23.00 |
| Fitness | 1 | 38 | 38.00 |
| Shopping | 1 | 6 | 6.00 |
| Medical | 1 | 84 | 84.00 |
| Overall | 27 | 798 | 28.44 |

Table 1: Distribution of blogs over categories. Columns **1**: Category; **2**: Number of blogs; **3**: Number of possessions identified in a category; **4**: Average number of possessions per blog in a given category.

## 5. Conclusion

In this article we introduced and defined the concept of possession identification, and shown how through the prism of the affective-cognitive consistency model, people's possessions may provide significant clues in inferring and analyzing people's behaviors and characteristics. We established an elaborate set of annotation guidelines, which enabled us to uncover 798 possessions from blog posts, with an agreement of 77%, indicating that the task is well defined. We are releasing the possession identification annotation guidelines, as well as three versions of the annotated blogs based on different confidence levels: gold-standard (where all three annotators agree), silver-standard (at least two annotators agree), and bronze-standard (containing all the annotations made), which we hope will kindle research into the area of possession identification.

## 6. Acknowledgements

This material is based in part upon work supported by National Science Foundation award #1344257, by grant #48503 from the John Templeton Foundation, and by DARPA-BAA-12-47 DEFT grant #12475008. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, the John Templeton Foundation, or the Defense Advanced Research Projects Agency.

## 7. Bibliographical References

Burger, J. D. and Henderson, J. C. (2006). An exploration of observable features related to blogger age. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 15–20.

Burger, J. D., Henderson, J., Kim, G., and Zarrella, G. (2011). Discriminating gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP 2011, pages 1301–1309.

[2] https://developer.walmartlabs.com/

[3] http://lit.eecs.umich.edu/research/downloads/

[4] http://www.nltk.org/api/nltk.tag.html

Cohen, R. and Ruths, D. (2013). Classifying political orientation on Twitter: It's not easy! In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM 2013)*.

Conover, M., Gonçalves, B., Ratkiewicz, J., Flammini, A., and Menczer, F. (2011). Predicting the political alignment of Twitter users. In *Proceedings of 3rd IEEE Conference on Social Computing (SocialCom)*.

Mukherjee, A. and Liu, B. (2010). Improving gender classification of blog authors. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 207–217.

Pennacchiotti, M. and Popescu, A.-M. (2011). Democrats, republicans and Starbucks afficinados: User classification in Twitter. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD 2011)*, pages 430–438.

Rao, D., Yarowsky, D., Shreevats, A., and Gupta, M. (2010). Classifying latent user attributes in Twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents*, SMUC '10, pages 37–44.

Rosenberg, M. J. (1956). Cognitive structure and attitudinal affect. *The Journal of Abnormal and Social Psychology*, 53:367–372.

Rosenberg, M. J. (1968). *Mathematical models of social behavior*, volume 1 of *The handbook of social psychology*. Addison-Wesley Pub. Co.

Van Durme, B. (2012). Streaming analysis of discourse participants. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 48–58.