

Сравнение ACE-разметки и разметки коллекций Persons-600/1000

Ниже приводится перевод фрагмента инструкции по разметке упоминаний лиц в соответствии с требованиями ACE. В примечаниях приводятся пояснения о содержательном различии ACE-разметки и разметки, выбранной для коллекций Persons-600 и Persons-1000.

Источник: http://projects.ldc.upenn.edu/ace/docs/English-Entities-Guidelines_v6.6.pdf

3.1 Персоны (PER)

Каждое отдельное упоминание персоны или группы лиц в документе считается ссылкой на сущность типа «Лицо». Например, упоминания людей могут выражаться как личное имя (“John Smith”), род занятий (“the butcher”), семейное положение (“dad”), местоимение (“he”) и т.д., а также какой-либо комбинацией перечисленного. Также к категории «Лицо» относятся трупы, человеческие останки, вымышленные персонажи-люди.

Метафорические упоминания также размечаются, если аннотатору очевидно, что это ссылка на человека.

*He is [a real turkey]¹
[The political cat of the year]
She's known as [the brain of the family]*

3.1.1 Подтипы персон

Лица делятся на следующие подтипы.

- PER.Individual — ссылка на отдельное лицо.
 - [Bill Clinton]
 - [The President of the U.S.]
 - The police found [[his] body]
- PER.Group — ссылка на группу лиц, не являющуюся организацией или GPE. Например, сюда относятся семьи, этнические и религиозные группы в отсутствие упоминания объединяющей их формальной организации.
 - [Analysts]
 - [IBM s lawyers]
 - [The family]
 - [The Kennedys]
 - [The house painters]
 - [The Arabs]
 - [the ethnic Albanians]
 - [The Christians]
- PER.Indeterminate — используется, если из контекста невозможно определить, идет ли отсылка к отдельному лицу или группе.
 - [You] want to go to the party.

¹ Квадратные скобки определяют протяженность именованной сущности, а подчеркивание будет обозначать главную часть (head).

Примечание [И.В.1]: В нашей разметке есть только упоминания в форме личных имен.

Дескрипции при личных именах «президент Буш» мы считаем отношениями между лицом и ролью лица (в данном случае, должностью) и не трактуем, как отдельное упоминание лица.

Лица, имена которых не выражены в тексте явно, например, «его отец», «президент США» также не размечаются по причине своего реляционного содержания.

Работу с неличными упоминаниями персоны, которая упоминалась в тексте в личной форме, мы склонны относить к задаче разрешения кореферентности и не считаем верным смешивать в одной коллекции два типа разметки. Выявление и извлечение именованных сущностей (лиц) — одна коллекция. Выявление всех упоминаний каждой именованной сущности — другая коллекция.

Примечание [И.В.2]: Мы не размечаем.

Примечание [И.В.3]: Мы размечаем только такой подтип.

- *If [you] believe in god, [you] must...*

3.1.2 Звания, обращения и должности

В английском большинство званий и обращений записываются перед упоминанием имени. Эти звания и обращения рассматриваются как **отдельные упоминания лица** (не рассматриваются как часть имени). Причем отдельные части звания или должности могут быть помечены **дополнительными тэгами**, если они ссылаются на другие именованные сущности. Например, в строке “US Secretary of State Madeline Albright” насчитывается четыре упоминания трех различных сущностей.

<i>[Secretary of State Madeline Albright]</i>	PER
<i>[Secretary of State]</i>	
<i>[US]</i>	GPE
<i>[State]</i>	ORG

Примечание [И.В.4]: Про дескрипции говорилось выше; мы не помечаем.

Примечание [И.В.5]: Мы в коллекции помечаем только один тип сущностей — лица.

Святые и другие личности, связанные с религиями

Религиозные саны (звания/должности), а также святые, пророки, имамы и архангелы трактуются как звания.

<i>[St. Christopher]</i>	PER
<i>[St.]</i>	

Примечание [И.В.6]: В нашей коллекции был бы помечен «Христофер» без слова «святной».

<i>[Pope John Paul II]</i>	PER
<i>[Pope]</i>	

Примечание [И.В.7]: В нашей коллекции был бы помечен «Иоанн Павел II» без слова «Папа».

Ссылка на «**Бога**» трактуется как имя.

<i>If you believe in [god], you must</i>	name mention
<i>Although he felt like he was [a god], he</i>	nominal mention

Примечание [И.В.8]: В нашей коллекции таких пометок нет.

«Пустые» названия должностей (“Empty” Job Titles)

Иногда названия должностей упоминаются нереферентно, как например в:

*She is **running for** President.*

В случаях, подобных этому мы не размечаем «President», т.к. это ссылка на должность, а не на лицо, ее занимающее.

Отметим, однако, что это не то же самое, что

Who will be [the next President]?

В этом случае мы помечаем [the next *President*] как PER.USP, т.к. это больше ссылка на неизвестное/гипотетическое лицо, чем на должность или пост.

Примечание [И.В.9]: Она баллотируется на пост президента.

Примечание [И.В.10]: Кто станет следующим президентом?

Примечание [И.В.11]: В нашей коллекции должности и посты не помечаются ни при каких обстоятельствах. Только личные имена.

3.1.3 Вымышленные персонажи, имена животных и имена вымышленных животных

Имена вымышленных персонажей помечаются, однако имена персонажей, использованный как названия TV-передач не помечаются, когда имеется в виду передача, а не персонаж.

[Batman] has become a popular icon
[Adam West] s costume from Batman the TV series

Имена животных не помечаются, так как они не ссылаются на персоны. То же касается вымышленных животных и персонажей-не-людей. Следующие два примера не содержат упоминаний лиц.

Morris the cat
Snuggle, the fabric softener bear

3.1.4 Части тел

Части тел помечаются **ТОЛЬКО** в случае, когда эта часть тела может в разумной степени определять человека, как в случае фрагменты трупа, собранные после боя, катастрофы и т.п.

Примечание [И.В.12]: В своей разметке мы придерживались этого принципа.

Примечание [И.В.13]: В нашей коллекции пометка будет для случаев, когда упомянуто личное имя — «останки Иванова».