Описание коллекции документов с пометками лиц, выраженных в тексте собственными именами

Сводная информация

Язык коллекции:	русский.
Формат документов:	плоский текстовый файл (plain-text).
Кодировка:	windows-1251.
Жанр:	новостные сообщения.

Назначение коллекции и формат разметки

Коллекция создана для оценки точности и полноты алгоритмов извлечения информации из текстов. Задача извлечения — выявление в тексте лиц, упомянутых в форме имен собственных. В рамках задачи предлагается определить места упоминания лиц в тексте, а также привести это упоминание к заданной *канонической* форме.

Каноническая форма имени представляется в виде строки следующего формата:

Имя␣{Индекс}␣Среднее имя␣Фамилия␣[уточнение]␣(оригинальное написание)

(в том числе отчество)

Здесь Индекс — это порядковый номер в династии: *Елизавета II*, *Генрих IV* и т.п.

Уточнение — это расширения фамилии в виде *Буш-отец*, *Кампард-старший* и т.п. Таким образом, в процессе структуризации извлекаемой информации предлагается отделить собственно фамилию от уточнения.

Оригинальное написание — это написание имени на языке-оригинале (обычно английском). Например,

В настоящее время ОВС НАТО командует адмирал Джеймс Ставридис (James G. Stavridis).

Все компоненты канонической формы разделены одним пробелом. Если в тексте какая-либо компонента отсутствует, то в канонической форме она пропускается.

Все компоненты в эталонной разметке записаны в верхнем регистре.

Если имя или среднее имя (отчество) упоминается в форме инициала, то в эталонной разметке присутствуют только буквы инициала (без точек). Но в компоненте *оригинального написания* точки после инициалов сохранены в эталонной разметке.

Приведем несколько примеров текстовых упоминаний лиц и соответствующих им канонических форм.

Текстовое упоминание	Каноническая форма
в телефонном разговоре с заместителем госсекретаря США Дэниэлом Фридом	ДЭНИЭЛ ФРИД
гендиректор компании McDonald's Дж.Скиннер	ДЖ СКИННЕР
Кампрад-старший больше не будет принимать участие	КАМПРАД [-СТАРШИЙ]
его сменит Мэрилин Хьюсон (Marillyn A. Hewson), которая	МЭРИЛИН ХЬЮСОН (MARILLYN A. HEWSON)
королева Елизавета II формально утвердила его в должности	ЕЛИЗАВЕТА {II}
поддерживать кандидатуру Джим Ен Кима при голосовании	ДЖИМ ЕН КИМ
срок полномочий В.Сердюкова истекал в июле	В СЕРДЮКОВ
Ж.-К.Трише получил пост главы ЕЦБ	Ж-К ТРИШЕ

При разметке не предполагалось установление каких-либо связей между различными упоминаниями одного лица. То есть если у нас в одном тексте есть упоминания *Борис Алешин* и *Б. Алешин*, то для каждого упоминания будет своя каноническая форма, а не наиболее полная информация.

Лица, для которых сегментация на элементы ФИО затруднительна (например, арабские имена, включающие в себя много атрибутивной информации и переводимые в духе «отец Халида и Фатимы из Багдада»; корейские имена, где фамилия упоминается перед именем), не структурировались и никаких перестановок компонентов имени для них не выполнялось. Но нормализация (приведение к канонической форме) выполнялась, если в русском языке традиционно принято склонение данного лица (*по словам Ким Чен Ира* → КИМ ЧЕН ИР).

Хранение коллекции в файловой системе

В файловой системе коллекция представляется как набор каталогов — по одному на каждый текстовый документ. В каждом каталоге хранится файл с текстом (text.txt) и файл с эталонной разметкой текста (anno.markup.xml).

Файлы с текстами — это плоские текстовые файлы в кодировке windows-1251.

Разметка хранится в xml-файлах следующей структуры. Корневой тэг именуется markup. В нем может быть произвольное количество элементов с тегом entry, каждая соответствует единичному упоминанию лица в тексте. Внутри тэгов entry могут располагаться следующие теги:

- id — не несет какой-либо смысловой нагрузки, но уникальный для всех entry в рамках одного файла anno.markup.xml.
- offset — смещение от начала плоского текстового файла (в символах), где начинается текущее упоминание лица.
- length — количество символов, использованное для упоминания лица.
- class — служебная константа, всегда имеет значение AAA_Estimate_Person.
- attribute — контейнер для пары тегов name/value. name всегда равно Canonical. Внутри тэга value хранится каноническая форма упоминания лица.

Пример элемента entry приведен ниже.

```
<entry>
<id>6</id>
<offset>2326</offset>
<length>14</length>
<class>AAA_Estimate_Person</class>
<attribute>
<name>Canonical</name>
<value>ТЬЕРРИ МУЛОНГЕ</value>
</attribute>
</entry>
<entry>
```

Особенности

Опишем некоторые «неоднозначные» конструкции, по которым трудно прийти к единому мнению, и принятые решения по их отражению в эталонной разметке.

Описание конструкции	Пример	Принятое решение по эталонной разметке
Уточнения журналистов, отделенные от основного текста.	*Как рассказал "Ъ" источник, знакомый с ситуацией, контракт господина Раппопорта продлен не будет. "Андрей Натанович (Раппопорт.- "Ъ") уходит, и вместо него пока будет назначен исполняющий обязанности",- сказал он.*	В разметку попали АНДРЕЙ НАТАНОВИЧ без фамилии и отдельно РАППОПОРТ без имени отчества.
	В пособничестве следователю обвиняют безработного Сергея Керимова (в некоторых источниках – Киримов),	В разметку попали СЕРГЕЙ КЕРИМОВ и отдельно КИРИМОВ.
Альтернативный вариант ФИО, указанный в скобках.	*Один из участников шпионского скандала в США в 2010 году - Майкл Зоттоли (Михаил Куцик) - получил должность в департаменте внешнеэкономической деятельности "Газпрома".*	МАЙКЛ ЗОТТОЛИ и МИХАИЛ КУЦИК размечены как отдельные эталоны.
Альтернативный вариант личного имени в скобках в составе ФИО.	*… состав нового правительства страны во главе с лидером победившей на выборах 1 октября 2012 г. коалиции "Грузинская мечта" миллиардером Бидзиной (Борисом) Иванишвили.*	В коллекцию такие примеры не включались.
	Эдуард Еунатович (Юрьевич) Худайнатов родился 11 сентября 1960 года в городе Чимкенте…	

Сочинительная конструкция с именами при общей фамилии.	*Первоначально ею управляли братья Дик и Мак Макдональды.*	Не размечалось.
Группировка лиц через дефис.	*По его словам, сохранится и "триумвират" Пейдж-Брин-Шмидт, который и будет принимать глобальные решения*	Каждый из топ-менеджеров Google размечен как отдельный эталон: ПЕЙДЖ, БРИН, ШМИДТ.
Лица, идентифицируемые собственным именем, не являющимся ФИО.	*он казался символом Москвы не менее вечным, чем ее святой покровитель Георгий Победоносец с копьем*	В коллекцию такие примеры не включались.
Упоминания личных имен в составе иных имен собственных (включая названия произведений литературы и искусства, транспортных средств, праздников, сооружений, телепередач и пр.).	*учился в Тихоокеанском высшем военно-морском училище имени С. О. Макарова, окончил Военно-морскую академию имени Адмирала Флота Советского Союза Н. Г. Кузнецова,*	Не размечалось.
	награжден орденом святого благоверного великого князя Дмитрия Донского первой степени.	
	… латиноамериканской прозы стал его роман "Смерть Артемио Круса" (1962).	
	… контролирующий сбор средств для издания доклада "Путин. Коррупция".	
Автонимные употребления имен	*… новый Папа выбрал для себя привычное в испаноязычном мире имя Франсиско — Франциск.*	Не размечалось.
	Новый Папа Римский принял имя Франциск I.	