Frequency Statistics of Words Used in Japanese Food Records of FoodLog

Sosuke Amano
Graduate School of
Interdisciplinary Information
Studies,
The University of Tokyo,
7-3-1 Hongo, Bunkyo-ku,
Tokyo 113-0033, JAPAN
s_amano@hal.t.u-tokyo.ac.jp

Makoto Ogawa
foo.log, Inc.
Entrepreneur Plaza 701,
7-3-1 Hongo, Bunkyo-ku
Tokyo 113-0033, JAPAN
ogawa@foo-log.co.jp

Kiyoharu Aizawa
Dept. of Information and
Communication Eng.,
The University of Tokyo,
7-3-1 Hongo, Bunkyo-ku,
Tokyo 113-0033, JAPAN
aizawa@hal.t.u-tokyo.ac.jp

Abstract
Recording foods enable us to improve our dietary habits. In food records, there are a variety of descriptions of meals because there is no standard way to express meal names. In this study, we analyze Japanese food records from the view of word frequency. We show very small numbers of words are satisfactory to describe the majority of the record.

Author Keywords
word frequency; food records; food log; power laws

ACM Classification Keywords
Categories and subject descriptors: H.2.8 [Database Management]: Database Applications---Data mining

Introduction
People are highly interested in food, and there are currently high expectations for information technology to contribute to improving diets. In particular, in the field of food records, a number of web services for health management and communication through dietary information have already been implemented [1, 2, 3]. However, it is difficult for non-experts to correctly record meals, and a definitive way to do this has not yet been established.
In this study, we investigated the frequency of words used in actual records of FoodLog [1], a food recording service developed for smartphones in Japan. Then, we categorized words that frequently appeared in the entire service and words that frequently appeared in data provided by specific individuals.

**Related Studies**

*Recipe Search*

In studies to discover rules for searching meal names, recipe data are mainly used [4, 7]. Recipe contains descriptions of ingredients and cooking methods for each meal, whereas food record data only deals meal names. The main difference between recipe and food record data is the appearance frequency of meal names. Since food records strongly reflect an individual’s meal preferences, a specific meal name appears many times. In contrast, it appears only once in recipe data.

**FoodLog Data**

The FoodLog food recording system [1] contains two databases (DBs) (Table 1): a DB containing 1,870 typical meal names that were prepared in advance (in-system DB) and a DB that is created and updated by each user (user-defined DB). The two databases are used in the following procedure.

1. Search for a meal name based on the system returns a list of keyword or an image candidates of the meal in the in-system DB and the user-defined DB.

2. Choose the corresponding one from the candidates.

3. When there is no proper name in the candidates, a new meal name is defined and registered to the user-defined DB. The meal name will be included in subsequent searches.

The user-defined DB is made and maintained for each user and is not shared. Therefore, a meal name defined by one user cannot be seen by another.

The data that were analyzed in this study were 457,944 records from 6,780 users created between May 31, 2013 and February 13, 2014. The number of records per user varied from 1 to 3,228.

Of the 1,870 meal names in the in-system DB, 1,842 meal names were used at least once. 77% of the total records were chosen from the in-system DB. In the user-defined DBs, there were a large number of meal names appeared; 34,224 meal names were registered.

<table>
<thead>
<tr>
<th>In System DB</th>
<th>ご飯(Rice), 味噌汁(わかめと豆腐) (Miso soup (seaweed and tofu)), ステーキ (牛サーロイン) (Steak (beef sirloin))</th>
</tr>
</thead>
<tbody>
<tr>
<td>User Defined DB</td>
<td>レーズンチョコ (Raisin chocolate), マクドナルド ボテト L (French fries L size from McDonald’s)</td>
</tr>
</tbody>
</table>

Table 1. Examples of meal names in the databases.

**Analysis of Frequently Appearing Words**

To analyze the meal name expressions in the FoodLog data, we segmented the meal names into words and extracted a set of frequently appearing words by calculating the frequency of each word. In addition, we analyzed tendencies that appeared by categorizing the extracted words.
Segmentation of a meal name into words
To segment a meal name into words, we used the Japanese morphological analysis library kytea [5], and the word boundary model was optimized by adding the top 2,000 most frequent meal names appearing in the FoodLog data to the general training data distributed in kytea.

Set of frequently appearing words
The appearance probability of a set of words \( W = \{w_1, \ldots, w_m\} \) in a food record data set \( D = \{d_1, \ldots, d_n\} \) is defined as

\[
P(W \mid D) = \frac{1}{S(D)} \sum_{w \in W} C(w \mid D),
\]

where \( S(D) \) is the sum of the number of appearances of all words in \( D \), and \( C(w \mid D) \) is the number of appearances of \( w \) in \( D \). In this case, the set of High-Frequency Words (HFW) in \( D \) is

\[
\text{HFW}(D) = \arg \min_{W} |W| = \left\{ W \mid P(W \mid D) \geq P_{th} \right\}
\]

HFW words appear more frequently than a rate of \( P_{th} \) in \( D \). In the analysis, we set \( P_{th} = 0.8 \).

Categorization of words
We categorized words into the following four types:

1. Ingredient or dish name, e.g., おにぎり (rice ball), パン (bread).
2. Cooking method, e.g., 焼き (grilled), 煮 (simmered), 生 (raw), 汁 (soup).
3. Food product name, e.g., ファミマ (Family Mart), SUBWAY.
4. Others, e.g., の (of), と (and), 健康 (healthy), グリーン (green), ml

In regard to errors in the word segmentation, we categorized words as type 1, 2, or 3 by correcting word segmentation only when that error occurred for the same word. When an error meant that there were multiple ways to classify a word, the word was categorized as type 4.

Food Record Analysis of FoodLog Data
Set of high-frequency words in the entire FoodLog data
We compare HFW of three user groups in FoodLog data, \( D_0, D_1, \) and \( D_2 \) (Table 2). The user groups contain different users who may have different eating habits. User grouping is automatic. The users were ranked by their amount of food records. According to the ranking order, every three users are categorized in a same group. We compute HFW(\( D_i \)) for each data set. We show HFW(\( D_0 \)) in Figure 1. Only 303 words are selected as HFW in \( D_0 \).

Following this, we calculated the appearance probability of HFW, \( P(\text{HFW} \mid D) \), for \( D_0, D_1, \) and \( D_2 \) (Table 3).
ヨーグルト ソフト ココア ジュース じゃがいも パイ 酔 きゅうり ちくわ 、 チョコ 餅類 スープ にんじん シチュー 発泡 さば 煮物 ライス
ご飯 かファラデ 海鮮 アボガド 割りレシピ パン 麺 みそ ミックス

肉 ( コーヒー ) 洗衣 かわせ 付き ナム ウィスキー ホット 煮物 ポテト カレー マカロニ 五目 切り たれ 、 ショートピザ さんま
玄 大根 大豆 飲み 梅干し 厚さ 高 6 グラス 酒 ビスケス 果 麦芽 カモナシ ゆで
納豆 味噌 アイス グレープ 麹 味 ちよう しもや 焼き ごはん ドリップ みかん フルーツ カップ せんべい フライド 一口 まなブラン パック 麻婆 ドレ 単位 ナッツ ロース しょうゆ えが とニンニク パン 丼

鶏肉 ( 味噌 ) 焼き エッグ 巨峰 西瓜 あし 千切り 鳥類 甘 ベーコン スムージー 低 たべ 焼き はみがく フレンチ まろ ようこそ にんじん オリジナル 肉 サラダ 食べ 方 にんじん オリガミ トースト 物 ウィン チョコレート 豆乳 冷やし 盛り 茶子
たこ 柿 あんかけ 麹物 ）合せ お ポーク チキン フライ にごりミート 醤油 すき さら クッキー ポンプ えび ミニ ささや さば 鯖 フライ ごはん

じゃが みる ほうれん草 ウインナーエム みそトースト に たまねぎ プレーン 、 かし する 生 メロン サラダ ( はるさめ ソテー 用 ラーメン ののり ぶどう キャベツ グリーン ロール 風 鉄板 パン ハン ブール うどん 中 天ぷら ソース おひたし チャーハン 餡料 小 オニオン みそ汁 オムレツ 醤油 トマト サラダ ゼリー ALCORE あん ピーマン 牛乳 ビーフ コロッケ 缶つ 血 トマト カフェオレ パンケーキ おでん スイス コーン あらし パスタ

 whey ワン切り おにぎり 切干し 田菜 タイプ 牛乳 ビール 米 カン パナナ おにぎり の カラメラ ナッツ トマト サラダ 醤油 お caves トースト チーズ ケーキ メロン 食べ 方 ケーキ 井 クリーム さつま 玉 かに ロース 魚 いか グラタン ソース インスタント トマト ステーキ 和え プラノ

Figure 1. The 303 words included in the set of frequently appearing words in the dataset D0. Words of a larger size appear more frequently. The top 10 most frequency words are "（”,” ”)，”ご飯 (rice)”，”焼き (roast)”，”サラダ (salad)”，”と (and)”，”汁 (soup)”，”味噌 (miso)”，”野菜 (vegetable)”

<table>
<thead>
<tr>
<th>DB</th>
<th>Users</th>
<th>Records</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>D0</td>
<td>2260</td>
<td>151473</td>
<td>430376</td>
</tr>
<tr>
<td>D1</td>
<td>2260</td>
<td>152635</td>
<td>432162</td>
</tr>
<tr>
<td>D2</td>
<td>2260</td>
<td>153836</td>
<td>437390</td>
</tr>
</tbody>
</table>

Table 2. FoodLog data of the three user groups.

<table>
<thead>
<tr>
<th>HFW</th>
<th>Size</th>
<th>D0</th>
<th>D1</th>
<th>D2</th>
</tr>
</thead>
<tbody>
<tr>
<td>HFW(D0)</td>
<td>303</td>
<td>0.8004</td>
<td>0.7843</td>
<td>0.7909</td>
</tr>
<tr>
<td>HFW(D1)</td>
<td>321</td>
<td>0.8033</td>
<td>0.8004</td>
<td>0.7980</td>
</tr>
<tr>
<td>HFW(D2)</td>
<td>311</td>
<td>0.7994</td>
<td>0.7890</td>
<td>0.8003</td>
</tr>
</tbody>
</table>

Table 3. Appearance probability of HFW in each data set.
Table 3 shows that the appearance probability is not very different from those defined by any data sets. We also see that only 300 words compose 80% of the words in the data sets, which includes 430,000 words and 150,000 food records. This is consistent with a power law [6] in that a small number of main elements make up the majority of the total. Note that the food record data is a mixture distribution of the frequency of meal name and that of words. 151,473 records in D0 included 14824 meal names, and they included 6186 words.

Frequently appearing words by users
To investigate the words recorded by an individual user, we defined User Particular Words (UPW) as

\[ UPW(D) = \arg \max P(W | D) = \{ W | |W| = 5, W \cap HFW = \emptyset \} \]

for the data set D, recorded by user i. User Particular Words are frequently appearing words for a specific user and are equivalent to the top 5 words that are not included in the HFW. We analyzed data sets for 954 users. Each of them had recorded more than 100 records (examples are shown in Table 4). We extracted 1,442 distinct words as the UPW. The words in both the HFW and UPW were categorized, and a comparison is shown in Table 5.

From the results in Table 4 and 5, in the UPW, which is a set of words that appear frequently only in some users, we see that there are more words related to company and product names such as Lawson and Nachuro (an abbreviation of Natural Lawson). There is no such feature in the HFW.

| User 1 | フルグラ (An abbreviation of fruit granola), 醋の (Vinegar), ローソン (Lawson), ナチュロー (An abbreviation of Natural Lawson), カルビ (Beef ribs) |
| User 2 | フレンチ (French), ハーゲンダッツ (Haagen Dazs), アンド (And), しょうゆ (Soy sauce), ケチャップ (Ketchup) |
| User 3 | カルビ (Beef ribs), ポトフ (pot-au-feu), サワー (sour cocktail), しらす (whitebait), だれ (sauce) |

Table 4. Examples of UPW defined by three users.

<table>
<thead>
<tr>
<th>Words</th>
<th>Dish name</th>
<th>Cooking method</th>
<th>Product</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>HFW(D0)</td>
<td>188</td>
<td>54</td>
<td>0</td>
<td>61</td>
</tr>
<tr>
<td>UPW</td>
<td>682</td>
<td>125</td>
<td>150</td>
<td>485</td>
</tr>
</tbody>
</table>

Table 5. Comparison of HFW and UPW categorization.

Summary
In this study, we have analyzed the present state of the Japanese food recording system, FoodLog, in terms of the frequency of words used in meal names. We have shown that most food records are composed of a small number of words, 300 words for composing 80% of 150,000 records. In addition, in records created by individual users, a meal is sometimes recorded using the company or product name. These results will need to be taken into account in future works.
Acknowledgements
This work was funded by CREST via JST and by a Grant-in-Aid for Scientific Research (B) from JSPS.

References