あと 1 文字で単語をなす文字集合について

浅原正幸　国立国語研究所　masayu-a@ninjal.ac.jp

概要

平仮名の多重集合 120 文字から 14 文字を集め、3 文字の単語を 4 つと 2 文字の単語を 1 つそろえる麻雀のようなゲームがある。本研究は、本ゲームを進めるにあたり必要な、あと 1 文字で単語をなす文字集合を構築した。その文字集合に基づくゲームの進行を、集合パッキング問題としてとらえ、配牌における和了（天和）・聴牌（ダブリー）確率などを計算した。結果、和了確率約 5.9%、聴牌確率約 40.8%、1 向聴確率約 40.5% であり、ゲームの本質が摸牌・打牌よりも理牌にあることを明らかにした。

勝つためにはどうすればいいか

★ここだけ読めばいい★

実質的には理牌を頑張るゲームである。理牌をがんばれば結構な確率で天和・地和が上がれる。

自然言語の文字の並びを集合ととらえたうえで、その文字があるかどうかを判定すればよい。作成した表 1_2chars.txt, 2_3chars.txt は、待ち候補文字が少ない順に並んでいる。待ち候補文字が多い塔子構成文字については、理牌時に最後に処理しても使える可能性が確率的に高い。一方、待ち候補文字が少ない塔子構成文字については、塔子に使いにくいため、待ち候補文字が少ない順に塔子を組むことが有利と考える。副露はしない。これが意識してできれば、シャンポン・ノベタンまでは考えなくてよい。

1　はじめに

本稿ではあるゲームについて考察するものである。本ゲームは、文字札を組み合わせて一定の構造を完成させることを目的とする競技型の思考ゲームであり、日本語の音韻体系を反映した文字集合を素材とする点に特徴がある。

まず本ゲームで使用される札について述べる。本ゲームでは、平仮名に相当する基本的な音節札を中心とし、それに濁音や半濁音、小さく書かれる文字、長音記号を加えた合計 120 枚の文字札を用いる[1)]。基本的な音節札の一部は同一文字が 2 枚ずつ存在し、その他の文字は 1 枚ずつである。

次にゲームの目的について記す。各プレイヤーは自分が保持する札を用いて、特定の規則に従う形を作り上げることを目指す。この形とは、3 つの文字が単語をなす組であったり、2 つの文字が単語をなす組であったりなど、一定の構造をもつ文字列の集まりである。これらの組を複数そろえることで、全体として完成した構造が成立する。

ゲームの進行は次のように行われる。プレイヤーは順番に山から札を 1 枚引き（摸牌）、不要な札を一枚捨てる（打牌）。この操作を繰り返しながら、手元の文字の並べ替えや選別を通じて完成形に近づけていく。他のプレイヤーが捨てた札を利用できる場合もあり、単に運だけでなく、状況判断や構造的予測が強く要求される。特に、あと 1 枚で完成が可能となる局面では、待っている文字の種類によっていくつかの形が区別される。本研究ではその部分局面の類型化として「あと 1 文字で単語をなす文字集合」のデータを構築する。

手札の状態は、完成までに必要な残り枚数（向聴数）によって段階的に評価することができる。すでに規則に従う形がすべて整っている場合は完成状態（和了）であり、あと一枚で完成となる場合はほぼ完成状態（聴牌＝向聴数 0）といえる。また、二枚以上必要な段階は途中状態（向聴数 1 以上）である。この「あと n 枚必要であるか」（n-1 向聴）という尺度によって、プレイヤーは自身の進行度を把握できる。ゲームの終了条件は、誰かが完成形を作り上げるか、山から札が尽きることである。本研究ではこの完成形および完成形に必要な残り枚数に対応する確率分布についても検討する。

2　あと 1 文字で単語をなす文字集合

整備した文字集合を https://github.com/masayu-a/hrgjncs に公開する。

1） この枚数はバリエーションがあるが、本研究では 120 枚ルールを適用する。

2.1 牌集合の仕様（120 牌）

本ゲームの内容物は次の通りで、総計は $\boxed{120}$ 牌である。

- 清音 45 音（「を」を除く）各 2 枚：計 90
- 濁音 20 音 各 1 枚：計 20
- 半濁音 5 音 各 1 枚：計 5
- 拗音 4 文字 各 1 枚：計 4（ゃ, ゅ, ょ, っ）
- 長音 1 文字 各 1 枚：計 1（ー）

これらに基づき、各種状態を付録の通り定義する。

2.2 妥当な単語集合の準備

ここでプレイヤーが単語と合意できる妥当な単語集合を準備する。単語親密度情報付き『分類語彙表』(WLSP-Familiarity) version 4.0[2] において「書く」の評定値が 0.0 以上である名詞（体の類）を対象とする。実際にはほかの品詞も含めることが可能であるが、ここでは配牌和了確率の下界を求めるために単語集合を制限する。

読みの部分を抽出し、カタカナ語などに対し適切な処理を行い、接尾辞相当の語（連濁）を排除するなどの手作業を行った。2 文字集合を 2gram.txt、3 文字集合を 3gram.txt とした。それぞれ 1 行目に可能な文字集合、2 行目に可能な単語（表記）を配した。

2.3 待ち形の定義

聴牌（$Shanten(H, M) = 0$) のとき，待ち集合 $Waits(H, M)$ の構造により，代表的な待ち形を以下のとおり分類する．本研究では，2gram.txt の 1 列目（2 文字集合）を B，3gram.txt の 1 列目（3 文字集合）を T とおく．また，辞書は「文字集合（ソート済み文字列）」で管理されるため，文字列 w をソートしたキーを

key(w) = (w の各文字を辞書順に並べ替えた文字列)

で表す．

待ち集合（基礎定義）

$$Wait_2(x) = \{\, y \in \mathcal{S} \mid \mathrm{key}(xy) \in B \,\}, \tag{1}$$

$$Wait_3(a, b) = \{\, c \in \mathcal{S} \mid \mathrm{key}(abc) \in T \,\}. \tag{2}$$

すなわち $Wait_2$ は「あと 1 文字で 2 文字単語」，$Wait_3$ は「あと 1 文字で 3 文字単語」を与える．

2） https://github.com/masayu-a/WLSP-Familiarity

2.3.1 2 文字単語完成待ち

定義（2 文字単語完成待ち）：

$$\begin{aligned} &\exists G_1, G_2, G_3, G_4 \in T,\ \exists x \in \mathcal{S} \\ \text{s.t.}\quad & H = G_1 \uplus G_2 \uplus G_3 \uplus G_4 \uplus \{x\}, \end{aligned}$$

このとき

$$Waits(H, M) = Wait_2(x).$$

すなわち，手牌はすでに 3 文字単語を 4 つ完成しており，残り 1 文字 x を 2 文字単語にするための「あと 1 文字」を待つ形である．

例：14 枚 (親配牌) の状態で

てがき / ねあげ / じもん / ろまん / ぬ / び

から ぬ を切って 13 枚にすると，び に対して $Wait_2$(び) が待ちとなる（例：え お か き く こ … 等，key(び y) $\in B$ を満たす文字）．

本研究 1_2chars.txt はこの単語集合を列挙する．

2.3.2 3 文字単語完成待ち

定義（3 文字単語完成待ち）：

$$\begin{aligned} &\exists G_1, G_2, G_3 \in T,\ \exists P \in B,\ \exists a, b \in \mathcal{S} \\ \text{s.t.}\quad & H = G_1 \uplus G_2 \uplus G_3 \uplus P \uplus \{a, b\}, \end{aligned} \tag{3}$$

このとき

$$Waits(H, M) = Wait_3(a, b).$$

すなわち，3 文字単語が 3 つと 2 文字単語が 1 つは完成しており，残り 2 文字 $\{a, b\}$ を 3 文字単語にする「あと 1 文字」を待つ形である．ここで $\{a, b\}$ 自体が 2 文字単語である必要はない（単なる未完成の 2 文字集合でもよい）．

2.3.3 2 文字単語から 3 文字単語完成待ち

定義（2 文字単語 →3 文字単語完成待ち）：

$$\begin{aligned} &\exists G_1, G_2, G_3 \in T,\ \exists P, Q \in B \\ \text{s.t.}\quad & H = G_1 \uplus G_2 \uplus G_3 \uplus P \uplus Q, \end{aligned} \tag{4}$$

かつ，Q が 3 文字単語へ拡張可能であるとする．このとき

$$Waits(H, M) = \{\, c \in \mathcal{S} \mid \mathrm{key}(Qc) \in T \,\}.$$

すなわち，13 枚時点で 2 文字単語 Q を完成として持っており，それにあと 1 文字を足して 3 文字単語へ“昇格”させることで和了形を満たす待ちである．（このとき P は最終的に 2 文字単語として残す役割を担う．）

本研究の 2_3chars.txt はこの単語集合を列挙する．

2.3.4 2 文字単語 2 つから 3 文字単語完成待ち

定義（シャンポン型）：

$$\exists G_1, G_2, G_3 \in T,\ \exists P, Q \in B \quad \text{s.t.} \quad H = G_1 \uplus G_2 \uplus G_3 \uplus P \uplus Q,$$

かつ，P と Q の双方が 3 文字単語へ拡張可能であるとする．このとき待ちは

$$Waits(H, M) = \{\, c \mid \mathrm{key}(Pc) \in T \,\} \cup \{\, c \mid \mathrm{key}(Qc) \in T \,\}.$$

すなわち，次の 2 通りの和了が（同一手牌から）成立しうる：

- P を Pc により 3 文字単語化し，Q を 2 文字単語として残す,
- Q を Qc により 3 文字単語化し，P を 2 文字単語として残す.

麻雀の二対子シャンポンに相当し，「どちらの 2 文字単語を 3 文字単語にするか」という二者択一の完成形が並立する待ちである.

例：14 枚の状態で

さろん / おやつ / ねむけ / こく / なに / て

から て を切って 13 枚にすると，こく を拡張して 3 文字単語にする待ちと，なに を拡張して 3 文字単語にする待ちが併存し，その和集合が $Waits(H, M)$ となる.

2.3.5 3 文字単語対の共有 2 文字による待ち

定義（ノベタン待ち）：

$$\exists G_1, G_2, G_3 \in T,\ \exists U, V \in T,\ \exists x, z \in \mathcal{S} \quad \text{s.t.} \quad H = G_1 \uplus G_2 \uplus G_3 \uplus C,$$

ただし C は 4 文字の多重集合で,

$$C = U \uplus \{x\} = V \uplus \{z\}, \quad U \neq V, \quad |U \cap V| = 2$$

（3 文字単語 U, V が 2 文字を共有し，合わせて 4 文字集合 C を成す）を満たすとする．このとき待ちは

$$Waits(H, M) = Wait_2(x) \ \cup\ Wait_2(z).$$

すなわち，4 文字部分 C から「どちらの 3 文字単語を採用するか」によって残り 1 文字が x または z となり，その残り 1 文字を 2 文字単語にするための待ちが 2 系統（2 行）併存する．本研究の `nobetan.txt` はこの 2 行を単位として列挙している.

3 和了・聴牌・向聴判定とその確率

本ゲームの和了形は「3 文字語を 4 つ＋ 2 文字語を 1 つ（雀頭相当）」である。公開済み語（副露に相当）を M、手牌を多重集合 H とし、辞書（2 文字集合）を B、（3 文字集合）を T とする。文字列 w の各文字を辞書順に並べ替えたものを $key(w)$ とし、辞書照合には $key(\cdot)$ を用いる。

3.1 判定アルゴリズム

向聴と待ち集合 向聴を $Shanten(H, M) \in \{-1, 0, 1, 2, \ldots\}$ とし、$Shanten = -1$ を和了、$Shanten = 0$ を聴牌とする。聴牌時の待ち集合は

$$Waits(H, M) = \{x \in \mathcal{S} \mid Shanten(H \uplus \{x\}, M) = -1\} \tag{5}$$

で与える。

集合パッキングとしての和了判定 手牌 H は多重集合であるため、実装上は各タイルにインデックス（札インスタンス）を付与して区別する。$|H| = n$ とし、札インスタンスの全集合を $U = \{1, \ldots, n\}$ とする。$X \subseteq U$ に対し、$H|_X$ を X が指す札からなる多重集合とする。

このとき、手牌上で作れる候補 3 語・候補 2 語を

$$\mathcal{G}_3(H) = \{X \subseteq U \mid |X| = 3,\ key(H|_X) \in T\}, \tag{6}$$

$$\mathcal{G}_2(H) = \{Y \subseteq U \mid |Y| = 2,\ key(H|_Y) \in B\} \tag{7}$$

と定義する。

副露済み文字の個数を $m = |M|$ とすると、和了は「$\mathcal{G}_3(H)$ から $4 - m$ 個、$\mathcal{G}_2(H)$ から 1 個を、互いに素になるように選んで U を完全被覆する」ことで判定できる。すなわち

$$\exists X_1, \ldots, X_{4-m} \in \mathcal{G}_3(H),\ \exists Y \in \mathcal{G}_2(H) \quad \text{s.t.} \begin{cases} X_p \cap X_q = \emptyset\ (p \neq q), \\ X_k \cap Y = \emptyset\ (\forall k), \\ X_1 \cup \cdots \cup X_{4-m} \cup Y = U. \end{cases} \tag{8}$$

これは「候補集合を互いに交わらないように詰め込む」集合パッキング（set packing）の形である。但し、以下では配牌時の和了判定を行うために $m = 0$ とする。

MIS（最大独立集合）への言い換え 候補集合（頂点）同士が札インスタンスを共有するときに辺を張った衝突グラフを考えると、(8) の「互いに素に選ぶ」は「独立集合として選ぶ」に一致する。よっ

Shanten	count	proportion	四麻 (最小; 七対子国士含)	四麻 (一般形のみ)	三麻 (2 萬–8 萬除外)
−1	1765	0.05883	3.02544×10^{-6}	2.67113×10^{-6}	9.45975×10^{-6}
0	12251	0.40837	0.000697837	0.000602410	0.00163934
1	12138	0.40460	0.023	0.020	0.041
2	3465	0.11550	0.195	0.167	0.251
3	380	0.01267	0.439	0.408	0.439
4	1	0.00003	0.285	0.299	0.240

表 1 14 枚配牌 (親) の向聴分布（本研究：シミュレーション推定，四麻・三麻：組合せ論的な厳密計数）

て本判定は MIS（最大独立集合）としても表現できる。

ビットマスク DP（$n = 14$ の小規模性の利用） 配牌は通常 $n = 14$ と小さいため、U を n ビットのマスクで表し、各候補 $X \in \mathcal{G}_3(H)$、$Y \in \mathcal{G}_2(H)$ を対応するビットマスク $mask(X), mask(Y)$ で管理する。和了判定は「3 語を $4 - m$ 個＋ 2 語を 1 個」選んだときに OR が全ビットになるか（完全被覆するか）で判定できる。向聴計算は「規定の語枠（3 語× $(4 - m)$ と 2 語× 1）で、手牌から何枚まで衝突なく説明できるか」の最大化問題として DP 化でき、最大被覆枚数を K とすると不足枚数は $(n - K)$ なので

$$Shanten = (n - K) - 1 \tag{9}$$

と置けば $K = n$ で $Shanten = -1$、$K = n - 1$ で $Shanten = 0$（聴牌）となる。

3.2 統計分析結果

14 枚配牌についてモンテカルロ・シミュレーションを行い，配牌時向聴 $Shanten$ の分布を推定した。具体的には，ランダムに生成した配牌を多数回サンプリングし，各配牌に対して判定アルゴリズム（§3.1）により向聴数を計算し，その度数分布を集計した（表 1 左は 30,000 試行の例）。その結果，配牌時和了（天和）（$Shanten = -1$）は 0.05883（約 5.9%），配牌時聴牌（ダブリー）（$Shanten = 0$）は 0.40837（約 40.8%）となり，さらに $Shanten = 1$ も 0.40460（約 40.5%）を占めた。平均向聴は概ね 0.6 程度であり，配牌時点で $Shanten = 0, 1$ が全体の大半を占めることが分かる。これは，本ゲームでは配牌の時点で完成形に近い手が比較的得られやすいことを示唆する。

本研究の確率値は上記のようにシミュレーションに基づく推定値である。これに対し，既存研究（例：Rascal のまとめ[3)]）では，${}_{136}C_{14}$（四人麻雀）や ${}_{108}C_{14}$（三人麻雀）等の総組合せ数を基礎として，向聴ごとの手牌数を組合せ論的に数え上げることで確率を与えている。すなわち，前者が「乱択サンプリングによる近似（推定）」であるのに対し，後者は「計数に基づく厳密計算」である。表 1 右には，四人麻雀・三人麻雀における 14 枚配牌時向聴分布（厳密計数）を，本論文の結果と同一の指標（向聴ごとの確率）で比較できるように整理して併記した。例えば四人麻雀では天和は 3.0×10^{-6} 程度，配牌時聴牌も 7.0×10^{-4} 程度に留まるのに対し，本ゲームではそれぞれ約 5.9%，約 40.8% と桁違いに大きい。以上より，本ゲームは一般的な麻雀に比べて配牌時点で和了・聴牌に到達している確率が高いことを示した。

4 おわりに

本研究では、平仮名の多重集合 120 文字から 14 文字を集め、3 文字の単語を 4 つと 2 文字の単語を 1 つそろえる麻雀のようなゲームを数理的に検討した。当初、14 文字の平仮名文字列を形態素解析すればよいと考えたが、最大 14!/5! [4)]通りの順列が考えられるために、これらを枚挙して形態素解析する方略は不適切と考えた。そこで、単語を文字集合ととらえたうえで、集合パッキング問題として和了判定ができると考えた。一般的に集合パッキングは NP 完全とされているが、本ゲームでは 14 枚固定で考えればよいために、現実的な時間で和了判定が可能であることを確認した。和了判定プログラムを拡張し、ランダムに発生させた配牌の向聴分布を計算し、本ゲームにおいては通常の麻雀よりも配牌時和了が高いことを示した。本研究では単語を名詞に限定したが、他の品詞も含めた場合はより配牌時和了が高くなることが推定される。

3） http://web.archive.org/web/20250331035354/http://www10.plala.or.jp/rascalhp/mjmath.htm

4） 14 文字配牌の順列を、$\{3, 3, 3, 3, 2\}$ の 5 単語の順列で割ったもの。

A　付録：ゲーム状態の定義

記号集合と多重度

記号集合 $\mathcal{S}$ を $\mathcal{S} = \mathcal{S}_{\mathrm{seion}} \cup \mathcal{S}_{\mathrm{dakuon}} \cup \mathcal{S}_{\mathrm{handakuon}} \cup \mathcal{S}_{\mathrm{yoon}} \cup \mathcal{S}_{\mathrm{choon}}$ とし、各部分集合を以下で与える（清音は「を」を除外）：

$$\mathcal{S}_{\mathrm{seion}} = \left\{ \begin{array}{cccccccccc} あ & い & う & え & お & か & き & く & け & こ \\ さ & し & す & せ & そ & た & ち & つ & て & と \\ な & に & ぬ & ね & の & は & ひ & ふ & へ & ほ \\ ま & み & む & め & も & や & & ゆ & & よ \\ ら & り & る & れ & ろ & わ & & & & ん \end{array} \right\},$$

$$\mathcal{S}_{\mathrm{dakuon}} = \left\{ \begin{array}{cccccccccc} が & ぎ & ぐ & げ & ご & ざ & じ & ず & ぜ & ぞ \\ だ & ぢ & づ & で & ど & ば & び & ぶ & べ & ぼ \end{array} \right\}.$$

$$\mathcal{S}_{\mathrm{handakuon}} = \{ ぱ \quad ぴ \quad ぷ \quad ぺ \quad ぽ \},$$
$$\mathcal{S}_{\mathrm{yoon}} = \{ ゃ \quad ゅ \quad ょ \quad っ \}, \mathcal{S}_{\mathrm{choon}} = \{ ー \}.$$

多重度関数 $\mu : \mathcal{S} \to \mathbb{N}$ を

$$\mu(x) = \begin{cases} 2, & x \in \mathcal{S}_{\mathrm{seion}}, \\ 1, & x \in \mathcal{S}_{\mathrm{dakuon}} \cup \mathcal{S}_{\mathrm{handakuon}} \cup \mathcal{S}_{\mathrm{yoon}} \cup \mathcal{S}_{\mathrm{choon}}. \end{cases}$$

と定義し、牌多重集合（全タイル）を

$$\mathcal{T}_{\mathrm{tiles}} = \biguplus_{x \in \mathcal{S}} \{x\}^{\mu(x)}$$

とする。総枚数は

$$|\mathcal{T}_{\mathrm{tiles}}| = 2 \cdot 45 + 20 + 5 + 4 + 1 = \boxed{120}.$$

局全体の状態

局全体 K　局の状態を

$$K = (S, \Sigma, \delta, s_0, F)$$

で与える。

状態集合 S　各状態を

$$s = (W, P, \mathrm{turn}, \mathrm{last}, \mathrm{phase})$$

とする。

- W：山（未使用の牌列；多重集合 $\mathcal{T}_{\mathrm{tiles}}$ の部分多重集合）
- $P = \langle p_0, p_1, p_2, p_3 \rangle$：4 人のプレイヤー状態。各 $p_i = (H_i, M_i, L_i)$
  - H_i：手牌（13 または 14 枚の多重集合）
  - M_i：副露済み文字（副露済み 3 文字単語の文字集合；麻雀のポン・チーに対応、配牌時は空集合とする）
  - $L_i \in \{0, 1\}$：リーチ状態フラグ
- $\mathrm{turn} \in \{0, 1, 2, 3\}$：手番プレイヤー ID
- last：直近の捨て札（出所プレイヤー ID と文字）または $\perp$
- phase $\in$ {Dealing, Draw, Discard, CallWindow, End}

入力記号 Σ **（イベント）**

- `draw`(i, c)：i が W から文字 c を自摸
- `discard`(i, c)：i が c を打牌（任意にリーチ宣言可能）
- `call`$(j, \mathrm{pattern})$：他家捨て字と自手で「語」を形成（チー／ポン相当）
- `ron`(j), `tsumo`(i)：和了
- `pass`(j)：鳴きパス

遷移関数 δ　イベントに応じて $W, P, \mathrm{turn}, \mathrm{last}, \mathrm{phase}$ を更新する（リーチ後の制約や見逃し規定を含む）。

初期状態 s_0　山順序と配牌（親 14・子 13）、起家、供託等を確定した状態。

受理状態 F　phase = End に到達した状態全体（和了／流局）。

和了形

麻雀における「4 面子 1 雀頭」に対応して、和了形（完成形）を

$$和了形 = 4\ 面子 + 1\ 雀頭$$

と定める。

語（3 文字）　`3gram.txt` にある 3 文字集合。

頭（2 文字）　`2gram.txt` にある 2 文字集合。

拗音・長音の扱い　$\mathcal{S}_{\mathrm{yoon}} = \{ ゃ, ゅ, ょ, っ \}$ および $\mathcal{S}_{\mathrm{choon}} = \{ ー \}$ は原則として単体で語・頭を構成しない。

向聴・聴牌の定義

手配 H と副露牌 M から必要追加文字数にもとづき定義する：

$$\mathit{Shanten}(H, M) \in \{-1, 0, 1, 2, \ldots\}.$$

$$\textbf{向聴数}:\ \mathit{Shanten}(H, M) = k, \tag{10}$$

$$\textbf{聴牌}:\ \mathit{Shanten}(H, M) = 0, \tag{11}$$

$$\textbf{和了}:\ \mathit{Shanten}(H, M) = -1. \tag{12}$$

聴牌時の待ち牌集合は

$$\mathit{Waits}(H, M) = \{ x \in \mathcal{S} \mid \mathit{Shanten}(H \uplus \{x\}, M) = -1 \}.$$

1 向聴（$\mathit{Shanten} = 1$）の改善集合は

$$\mathit{Improve}_1(H, M) = \{ x \in \mathcal{S} \mid \mathit{Shanten}(H \uplus \{x\}, M) = 0 \}.$$

状態遷移（要約）

- `draw`(i, c)：14 枚化 → $\mathit{Shanten}$ 再計算。-1 なら `tsumo`(i) で終了。
- `discard`(i, c)：13 枚化。最善打なら $\min_{c \in H_i} \mathit{Shanten}(H_i \setminus \{c\}, M_i)$。
- `call`$(j, \mathrm{pattern})$：捨て字を用いて語を形成し M_j に追加、$\mathit{Shanten}$ 更新。
- `ron/tsumo`：和了成立で End。