トピックモデルを用いた益城町仮設住宅聞き取り調査の自由回答の基礎分析

川野　倫輝[1]・佐藤　嘉洋[2]・円山 琢也[3]

[1]学生会員　熊本大学大学院自然科学研究科社会環境工学専攻（〒860-8555 熊本県熊本市中央区黒髪 2-39-1）
E-mail:178d8811@st.kumamoto-u..ac.jp

[2]学生会員　熊本大学大学院自然科学研究科社会環境工学専攻（〒860-8555 熊本県熊本市中央区黒髪 2-39-1）
E-mail:yo-sato@kumamoto-u.ac.jp

[3]正会員　熊本大学准教授　くまもと水循環・減災研究教育センター
（〒860-8555 熊本県熊本市中央区黒髪 2-39-1）
E-mail:takumaru@kumamoto-u.ac.jp

本研究は，仮設住宅居住者を対象とした聞き取り調査の自由回答に，自然言語解析の一手法であるトピックモデルを適用した分析を報告する．トピックとして話される確率が高いものに「自宅の解体」や「交通の不便さ」であることがわかった．また，対話時間推定モデルと組み合わせた分析より，「仮設住宅周辺のバリアフリー」，「経済的な負担と行政からの補助」，「情報提供の方法」，「仮設後の住まい」，「道路拡幅などの復興まちづくり」，「仮設団地内の設備要望」，「コミュニティ」などのトピックを話す回答者の対話時間が長くなるという結果が得られた．

Key Words: *2016 Kumamoto earthquake, temporary housing, topic model, duration model*

1. はじめに

2016 年 4 月の熊本地震が発生して以降，被災前の生活を取り戻すことのできない被災者は依然として多く存在する．このような被災者の実態把握や早期復興へ向けた知見を得るための調査が行政や地域支えあいセンターなどによって継続的に行われている．多くの社会調査において自由回答項目が設けられるものの，一般的に，十分に集計・分析がなされているとは言えない．2016 年 6 月から熊本大学が実施した益城町仮設住宅聞き取り調査においても，基礎集計こそ益城町復興計画の資料 [1]にも示されていたが，自由回答の分析は行われていなかった．

この点を問題意識として，筆者らは本調査の自由回答分析 [2]を行った．対話時間と自由回答中の異なり語数を記述する数理モデルの適用から，高齢者が対話中に同じ話題の話を繰り返す等の示唆を得ている．また，対応分析により，属性と用いられた語の傾向を把握している．しかし，ここにおける話題とは，対話時間と異なり語数の関係から考慮したものである．具体的には，対話時間が長くなるにもかかわらず異なり語数が少ないという場合，同じ語が集中的に用いられたと考え，話題少なくなるという分析である．聞き取り調査において，調査員との対話時間は，回答者の本音や悩みを聞き出せているかの重要な指標となりうるためその分析は重要である．しかし，この分析では語の同義関係や共起性は考慮できておらず，話題について直接的に触れることはできてはいない．

そこで，本研究では，トピックモデルを用いた自由回答分析を行う．トピックモデルとは，近年，自然言語処理の分野で発達してきた文書生成モデルである．潜在的意味解釈により，文書中のトピックの抽出などに利用されており，土木計画学における適用研究 [3]も存在する．

本研究では，トピックモデルとして Latent Dirichlet Allocation (LDA) を用いた．LDA は一文書中に複数のトピックが存在することを許容しており，被災時の経験から復興計画への意見まで幅広く話を伺っている益城町仮設聞き取り調査への適用性は高いと考えられる．本研究では，LDA を利用し益城町仮設聞き取り調査の自由回答からトピックの抽出を行い，その推移を確認する．そして，数理モデルを適用した分析から対話時間とトピックの関係を考察する．

表-1 仮設住宅以降のお住まいについての意識調査の調査内容

	調査項目
問 1	震災前の住所
問 2 (1) (2) (3) (4)	震災前の住まいについて 住宅の所有 居住年数 住宅の形態 居住スペース以外の用途
問 3	自宅の被災状況
問 4	仮設住宅後の住まいの希望
問 5 (1) (2) (3)	家族について 現在の世帯構成と自動車保有台数 普段の生活で最もよく行くところ 震災前との世帯人数の変化
問 6 (1) (2)	益城町の将来について 益城町の復興・復旧において重要と思う点 益城町の復興計画を作るにあたっての意見や要望
問 7	行政，大学などへの意見・要望も含めて，現在の気持ち・心境

2. 益城町仮設住宅聞き取り調査

(1) 益城町仮設住宅聞き取り調査の概要

分析対象は先述の益城町仮設住宅聞き取り調査である．本調査は，益城町の仮設住宅居住者を対象に，訪問面接調査として行われた．調査目的は以下の2点である．

- 町民が必要とする災害公営住宅の戸数，希望する場所などの把握
- 現時点で不自由な点，不安などの幅広い把握

調査内容の概要は**表-1**に示す．このうち，本研究において分析に利用する自由回答形式の設問は，「益城町の復興計画を作るにあたっての意見や要望」と「行政，大学などへの意見・要望も含めて，現在の気持ち・心境」である．その他にも，設問の答えに該当しない回答者の発言も極力全て自由回答として記入しており，これも分析の対象となっている．しかし，回答者の発言を自由回答として取り上げ，記入するか否かは調査員に依存する部分がある点には留意が必要である．

本調査は，2016年6月30日から同年11月20日までに行われた．この期間内に，1,196戸の調査を完了した．これは分析対象の17団地中，未入居世帯を除いて81.4%の実施率となる．なお，調査票や基礎集計などは益城町復興計画の資料[1]や渡邉らの研究[4]を参考されたい．

(2) 自由回答有無の定義

ここで，自由回答ありとみなす調査票について説明をする．自由回答に相当する話が無かった調査では，調査員が「特になし」と調査票に記入している場合がある．このような調査票は自由回答なしとして扱われるべきものであるので，自由回答が以下の文章のみであったものを除くこととした．

- 特になし
- 今は特になし
- 特に困っていることはない

この処理を経ると，自由回答があるとみなす調査票は1,087戸分となった．これは分析対象中90.9%である．

3. トピックモデルの適用

(1) 事前処理

a) 形態素解析

語数を算出するために形態素解析を行う．形態素解析器として「MeCab」を用いる．形態素解析とは，文章を意味のある単語に区切り，辞書を利用して品詞や内容を判別することである．例えば，「解体を早くしたい」であれば，「解体」「を」「早い」「する」「たい」のように5つの語に区切ることができる．

b) 語数の集計結果

形態素解析の結果，総語数は55,851語，総語中の異なり語は4,102語であった．異なり語とは，同一の単語が何度用いられていてもこれを一語とし，対象の文章中に異なる単語がいくつあるかをかぞえた数である． これに対して，総抽出語数を延べ語数と呼ぶこともできる．このうち，分析に利用するのは名詞，形容詞，動詞(非自立語，接尾語，数を除く)のみである．この処理を経ると，総語数は21,511，異なり語数は3,591語となった．

c) 語の特徴量

文書内の語の特徴量の尺度として tf-idf 値を用いる．tf 値(ターム頻度)は一文書内での語の出現しやすさを示し，idf 値(文書頻度の対数)は語が特定の文章中で集中して用いられることを表す．よって，これらの積で表される tf-idf 値は，その値が大きい語ほど文書内での重要度が高いことを示す．以下に ti-idf 値の算出法を示す．

$$\text{tf-idf}_{kv} = \frac{ni,j}{\Sigma knk,j} * \log \frac{|D|}{dfi} \quad (1)$$

n_{ij} : 単語 i の文書 j における出現回数

k : 各文書中の語数

D : 総文書数

後述の Bag of Words(BoW)では，この値を語の重みとし，出現回数との積をとることとする．

(2) Latent Dirichlet Allocation(LDA)の概要

本研究では，トピックモデルとして，Latent Dirichlet Allocation(LDA)を利用する．ここで，の概要について説明する．なお，LDA についての解説は参考文献[4][5]に基づいて示す．

LDA はBag of Words(BoW)表現された文書集合を生成するための確率モデルである．BoW表現とは，文章中に

現れる単語のベクトル表現である．また，BoW表現は文章の構造は無視しており，単語の出現回数と共起性を表している．LDAは，BoWから得られる単語の共起性を用いて単語や文書をクラスタリングする手法として用いられる．

LDA では，文書中の各単語に，BoWからは直接得ることのできない潜在変数（トピック）を仮定する．また，LDAの特徴として，文書は複数のトピックから構成され，トピックの構成比としての確率分布をもつ．具体的には，文書d のi 番目の単語をw_{di} として，対応する潜在変数をz_{di} と定義する．ここで，トピック数をKとし，$\theta_{dk}(k=1,2,\ldots,K)$を文章$d$ でトピックk が出現する確率とする．トピック分布は$\theta_d=(\theta_{d1}\ldots,\theta_{dK})$ となる．また，各トピックはそれぞれに対応した単語の出現分布$\phi_k(k=1,2,\ldots,K)$ を有している．文書数をD，文書d の文章長（総単語数）をN_dとする． ϕ_{dv} をトピックk における単語v の出現確率とし，単語の出現分布を$\phi_k=(\phi_{k1}\ldots,\phi_{kV})$ とする．

θ_dやϕ_{kl}はDirichlet分布による生成を仮定すると，

$$\theta_d \sim \mathrm{Dir}(\alpha), d=1,\ldots,M \quad (2)$$

$$\phi_k \sim \mathrm{Dir}(\beta), k=1,\ldots,K \quad (3)$$

ここで，ハイパーパラメータα，βはそれぞれトピック数K，単語数Vの次元をもつ．

以下に，LDAにおける文書生成過程の流れを示す．

1. 文書毎に，事前分布のディレクレ分布 $\mathrm{Dir}(\alpha)$ に従い，トピックの出現確率分布θ_{dk} を生成する．
2. トピック毎に，事前分布のディレクレ分布 $\mathrm{Dir}(\beta)$ に従い，単語出現確率分布 ϕ_kを生成する．
3. 文書内の単語毎に，1.で生成したトピックの出現確率分布θ_{dk} に従い，トピック z_{di} を生成する．
4. 文書内の単語毎に，2.で生成したトピックの単語出現確率分布ϕ_k に従い，単語 w_{di}を生成する．

図-1にLDAの文書生成過程を表すグラフィカルモデルを示す．

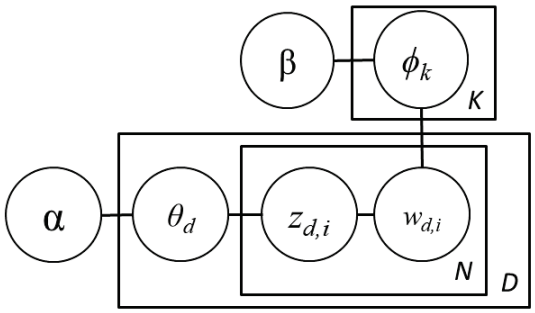

図-1　LDAのグラフィカルモデル

(3)　トピックの抽出結果

サンプリング方法には，崩壊型ギブスサンプリングを用いた．ハイパーパラメータは既往研究[6]を参考にα=1.0,β=1.0とした．サンプリング数は1000回とし，トピック数は15とした．**表-2**に抽出された各トピックの上位20語とトピックタイトルを示す．これは，**表-1**に示す単語をもとに，これらの文章中での用いられ方も考慮してトピックのタイトル付けを行ったものである．具体的には，「避難生活」「仮設住宅周辺のバリアフリー」「交通の不便さ」「家族の様子」「経済的な負担と行政からの補助」「騒音やプライバシー」「情報提供の方法」「仮設後の住まい」「道路拡幅などの復興まちづくり」

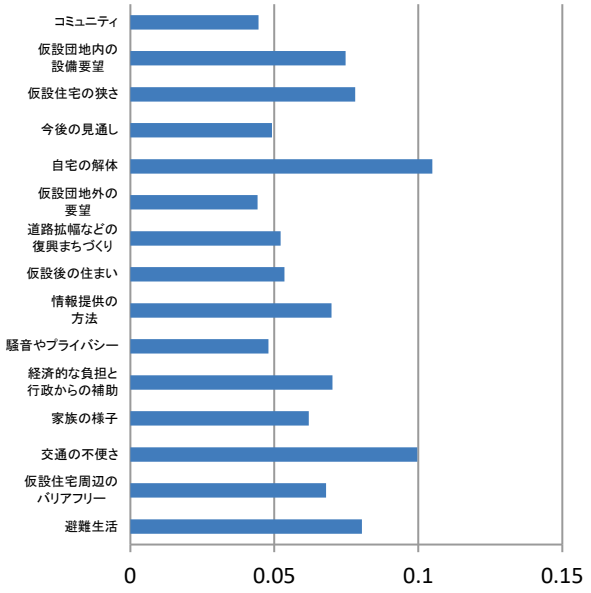

図-2　トピック割合の平均値

表-2　抽出されたトピック

避難生活	仮設住宅周辺のバリアフリー	交通の不便さ	家族の様子	経済的な負担と行政からの補助	騒音やプライバシー	情報提供の方法	仮設後の住まい	道路拡幅などの復興まちづくり	仮設団地外の要望	自宅の解体	今後の見通し	仮設住宅の狭さ	仮設団地内の設備要望	コミュニティ
知り合い	段差	バス	自宅	行政	気	情報	住宅	災害	ペット	建てる	先	狭い	ほしい	コミュニティ
体育館	手すり	買い物	暮らす	何	騒音	入る	益城	車線	益城町	早い	立つ	置く	近く	田んぼ
いる	トイレ	車	実家	時間	部屋	掲示板	復興	まちづくり	拡幅	土地	見通し	部屋	街灯	みんな
地震	風呂	移動	入居	町	音	分かる	気持ち	県道	高齢者	公費	冷蔵庫	荷物	暗い	集まる
娘	前	バス停	息子	かかる	落ちる	避難所	防災	強い	楽しみ	解体	住む	駐車	ポスト	相談
避難所	言う	便	全壊	ゴミ	瓦礫	人	不満	情報提供	それ	家	熊本市	置ける	いう	把握
総合	持つ	不便	夫婦	業者	釘	作る	状態	道路	地盤	急ぐ	仮設住宅	スペース	ない	誰
仕事	玄関	運転	小屋	お金	処理	良い	基準	自営	道路	元	やる	仮設住宅	コンビニ	飯野
仮設	冬	タクシー	出る	負担	宮園	町	2年後	町	若者	戻る	声	生活	団地	活動
車中泊	高い	交通	半壊	主人	ドア	回覧板	ない	体制	整備	再建	問題	付き合い	自転車	教育
避難	比べる	危ない	犬	仕事	車	集会	思う	瓦礫	支援	地区	農家	収納	子供	やる
鍵	つける	遠い	子供	解体	うるさい	聞く	住める	困難	こちら	進める	態度	ある	市内	年齢
息子	使う	悪い	暮らし	人	分別	知る	出来る	残す	重要	道路	2年間	暑い	入り口	自治会
行く	干す	便利	前	ない	プライバシー	近所	面	方法	外	自分	困る	仮設	施設	杉堂
修理	スロープ	今	慣れる	心配	周辺	安全	これ	高森線	示す	いつ	目途	復旧	テクノ	わかる
それ	悪い	一人	生活	他	助かる	なる	配慮	期間	街	作業	自分	近所	遠い	移転
年金	足	病院	記入	しない	危険	ルール	町営	事務所	広い	予定	駐車場	必要	公園	健康
手続き	もらう	送迎	夫	変わる	雨	どこ	子供	情報	子	心配	子供達	せまい	場所	環境
予定	生える	団地	本震	借りる	仕方	人達	被災	昼間	活性	できる	若い人	ほしい	ATM	暑い
以前	大変	嬉しい	仮説	離れる	街灯	インターネット	聞く	処理	必要	撤去	倒壊	いる	買い物	つくる

「仮設団地外の要望」「自宅の解体」「今後の見通し」「仮設住宅の狭さ」「仮設団地内の設備要望」「コミュニティ」となっている．「避難生活」は震災後，避難所として体育館に寝泊まりしてい**図-2**にはトピック割合の平均を示した．「自宅の解体」が最も高くなっており，10%を超えている．次いで割合が高くなっているのは「交通の不便さ」となっている．

図-3に性別でのトピック割合を示す．性別にかかわらず，トピック割合が高くなっているのは，「交通の不便さ」「自宅の解体」「仮設の狭さ」である．このほかに男性で割合が高いのは，「経済的な負担と行政からの補助」「仮設団地外の要望」であり，女性では「情報提供の方法」「仮設団地内の設置要望」の割合が高くなっている．「仮設団地外の要望」とは，地盤の調査や町内会などの仮設団地の外のものごとに関する要望であり，「仮設団地内の設置要望」とは，街灯やポスト，ATMなど現在または団地内という比較的近い範囲での要望に関するトピックである．また，「避難生活」や「家族の様子」も女性のほうがトピック間順位が高い．これらから，男性では，経済的な事や行政に関するトピックが選ばれる一方で女性では身近な生活への関心が強く，自分や身近な人についての話をする傾向があると考えられる．

図-4，**図-5**に年齢別で見たトピック割合を示す．トピック数が多く，図が煩雑になるため図は2つに分割した．

高齢者ほど割合が高くなるものとして，「避難生活」「仮設住宅周辺のバリアフリー」「情報提供の方法」がある．高齢者は情報の獲得に困難を抱えており，回覧板や掲示板を用いた情報提供を望んでいると推測される．若年層の割合が高くなるものには「騒音やプライバシー」「自宅の解体」があった．割合の分布が上に凸の形状を取るものに，「家族の様子」「経済的な負担と行政からの補助」「今後の見通し」「仮設住宅の狭さ」「仮設団地外の要望」「仮設の狭さ」があった．このうちは，「経済的な負担と行政からの補助」「今後の見通し」は40歳代～70歳代の間でピークを取っている．この年齢層は世帯主の多い年齢層と予想される．これらは世帯の意思決定に関わるトピックであるため割合が高くなっていると考えられる．一方，下に凸の形状を取るものに，「交通の不便さ」「仮設後の住まい」「仮設住宅内の設備要望」「コミュニティ」がある．

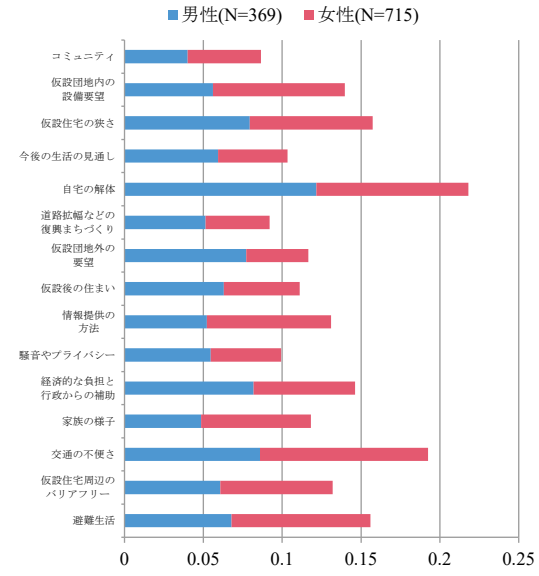

図-3　性別のトピック割合

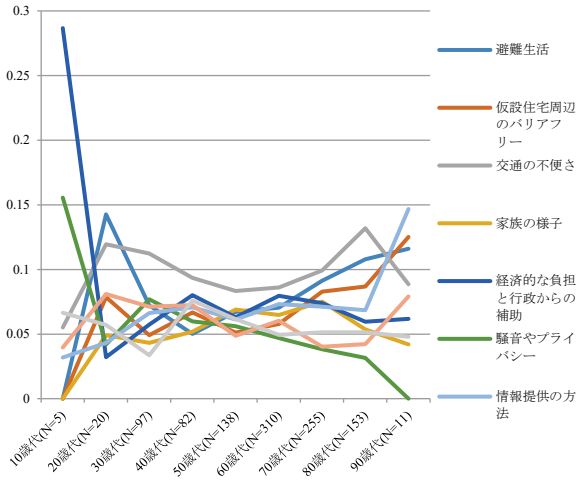

図-4　年齢別のトピック割合

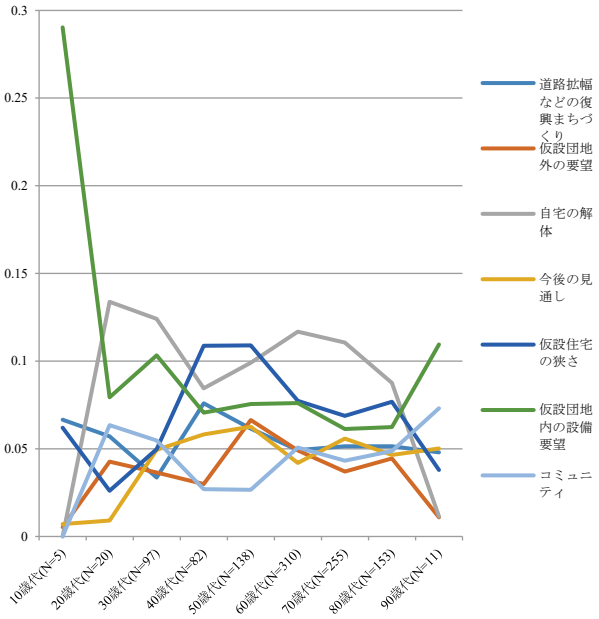

図-5　年齢別のトピック割合

4.　対話時間の推定

(1)　対話時間モデル

前章で抽出したトピックを用いて，対話時間の推定を行う．ワイブル分布を仮定した生存関数を利用し，対話時間モデルの構築をする．生存関数とは，生存状態から時間 t が経過した後に死亡状態へ移行する割合を算出す

る関数がある．これより，生存状態を調査が行われている状態，死亡状態を調査が終わっている状態と考えると，調査を開始した時刻から時間 t の経過と共に終了した調査件数を表現できる．

時点 t において個人 n が調査を行っている確率 $S(t)$ は以下になる．

$$S(t) = \exp(-\lambda t^{\alpha}) \quad (4)$$

よって，ある時点 t において個体 n が調査が終わっている確率 $F(t)$ は次式で表わされる．

$$F(t) = 1 - \exp(-\lambda t^{\alpha}) \quad (5)$$

また，その確率密度関数は次式で表わされる．

$$f(t) = \frac{dF(t)}{dt} = \alpha\lambda t^{\alpha-1} \cdot \exp(-\lambda t^{\alpha}) \quad (6)$$

ここで，パラメータ α はその時間的スケールに対応した尺度パラメータであり，この値が大きくなるほど対話時間が短くなることを示す．パラメータ λ は分布の形状を決める形状パラメータであり，この値が小さいほど，早い時間に急激に調査を終了する確率が高くなることを示す．一般的な生存関数では指数分布を用いるが，本研究では分布の形状を決定する形状パラメータを考慮することができ，対話時間を詳細に表現できるワイブル分布を仮定する．α>1の場合，ハザード比は時間の経過と共に増加する．一方，α<1の場合ハザード比は時間の経過と共に減少する．α=1の場合は指数分布で，ハザード比は一定である．

本研究では，上記のパラメータ λ を以下のように定式化した．

$$\lambda = \exp(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n) \quad (7)$$

ここで，x_i はトピック分布や個人属性などの説明変数である．

(2) 対話時間モデルの推定結果

本研究では，自由回答のあるデータのうち各個人属性の欠損のない1057サンプルにおける対話時間モデルを推定した．分布の形状を決める形状パラメータには，説明変数として農家ダミー，非就業者(65歳以上)ダミー，年齢，4章で抽出したトピック割合と各回答者の用いた総語数の積，定数項を入れている．トピック割合と各回答者毎の総語数の積は，各トピック毎の期待値を表す．パラメータが正であるほど調査を打ち切りやすい（対話時間が短くなる）ことを意味している．形状パラメータは小さい値になるほど早期に急激に調査を終了する確率が高くなることを意味する．表-5の推定結果より，すべての変数が有意となった．個人属性でみると，高齢者や農家，65歳以上の非就業者で調査時間が長くなる傾向にあることがわかる．トピックでみると，「仮設住宅周辺のバリアフリー」，「経済的な負担と行政からの補助」，「情報提供の方法」，「仮設後の住まい」，「道路拡幅などの復興まちづくり」，「仮設団地内の設備要望」，「コミュニティ」の期待値が高いほど調査時間が長くなる傾向にあることがわかる．ここから，これらのトピックが調査回答者の本音や悩みとなる可能性が高いと考えられる．

6. 結論

本研究の成果を以下にまとめる．

- トピックモデルより，トピック数を15と設定すると，「避難生活」「仮設住宅周辺のバリアフリー」「交通の不便さ」「家族の様子」「経済的な負担と行政からの補助」「騒音やプライバシー」「情報提供の方法」「仮設後の住まい」「道路拡幅などの復興まちづくり」「仮設団地外の要望」「自宅の解体」「今後の見通し」「仮設住宅の狭さ」「仮設団地内の設備要望」「コミュニティ」のトピックが得られた．
- トピックとして話される確率が高いものは「自宅の解体」や「交通の不便さ」である．
- 「仮設住宅周辺のバリアフリー」，「経済的な負担と行政からの補助」，「情報提供の方法」，

表-3 対話時間推定モデルの推定結果

説明変数		推定値	t値
形状パラメータ λ	仮設住宅周辺のバリアフリー	-0.02	-3.48
	経済的な負担と行政からの補助	-0.01	-3.18
	情報提供の方法	-0.02	-3.81
	仮設後の住まい	-0.01	-3.02
	道路拡幅などの復興まちづくり	-0.02	-2.70
	仮設団地内の設備要望	-0.01	-3.32
	コミュニティ	-0.02	-3.27
	農家ダミー	-0.22	-3.55
	65歳以上非就業者ダミー	-0.11	-2.27
	年齢	-0.01	-7.54
	定数項	-1.97	-23.62
最大対数尤度		-3677	
サンプル数		1057	

「仮設後の住まい」，「道路拡幅などの復興まちづくり」，「仮設団地内の設備要望」，「コミュニティ」は被災者の本音の悩みとして考えられる．

今後の課題としては適切なトピック数の決定が挙げられる．トピックモデルの結果はあらかじめ設定するトピック数に依存する．階層ディリクレ過程を用いる，またはトピック毎の類似度などを考慮したトピック数の適切な設定を行う必要がある．

今回は震災復興初期のデータを対象とした分析であったが，今後の展開としては，他時点の調査データを用いた時点間の比較が挙げられる．またトピックは何らかの構造を有していると考えられる．LDAの拡張である相関トピックモデルや階層構造を導入したトピックモデルを適用することも価値が高いと考える．

謝辞：仮設住宅聞き取り調査は，益城町復興課との共同実施によるものです．また青山学院大，慶応義塾大，東京大，自治医科大，関西学院大，京都大，九州大，九州工業大，佐賀大，大分大，鹿児島大，熊本学園大，熊本県立大及び熊本大の学内の多くの皆様にボランティアでご協力をいただきました．深く感謝申し上げます．

参考文献

1) 熊本県益城町: 益城町復興計画, pp. 99-111, 2016.12
2) 川野倫輝，佐藤嘉洋，円山琢也：益城町仮設住宅聞き取り調査の自由回答分析，第 55 回土木計画学研究発表会(春大会)，2017.6.
3) 塚井誠人，椎野創介：討議録に対するトピックモデルの適用，土木学会論文集 D3,Vol.72,No.5, pp. I_341-I_352, 2016.
4) 渡邉萌，佐藤嘉洋，円山琢也：熊本地震における益城町仮設住宅入居者の居住地選択意向分析，第 55 回土木計画学研究発表会(春大会)，2017.6.
5) Blei, D. M., Andrew, Y. N. and Michael I. J.: Latent dirichlet allocation, *Journal of Machine Learning Research* 3, pp.993-1022, 2003.
6) 岩田具治：トピックモデル，講談社，2015.
7) 加藤嘉浩：Latent Dirichlet Allocationの漸近解析，電気通信大学修士論文，pp.21-41,2013.

(2017.7.31 受付)

ANALYZING FREE ANSWER OF INTERVIEW SURVEY AT TEMPORARY HOUSING IN MASHIKI TOWN, KUMAMOTO USING TOPIC MODEL

Tomoki KAWANO, Yoshihiro SATO and Takuya MARUYAMA

This study analyze free answer of interview survey for temporary housing residents using a topic model that is one method of natural language analysis. We found that "home dismantling" and "travel inconvenience" are two topics with high spoken probability . Duration model analysis of interview time revealed that topics brings longer interview time are; "barrier-free around temporary housing", "financial burden and support from goverment", "information provision ", "post-temporary housing", "reconstruction planing including road widening", "equipment requests in temporary housing ", and "community".