ニューラル言語モデルの学習初期における単語の分節化

帖佐 宗浩　西田 悠人　大羽 未悠　渡辺 太郎
奈良先端科学技術大学院大学
chosa.munehiro.cl6@naist.ac.jp
{nishida.yuto.nu8, oba.miyu.ol2, taro}@is.naist.jp

概要

人間の乳幼児はその言語獲得の初期段階において，複数の単語からなる系列を，まとめて「ひとかたまり」で意味をなす系列（ホロフレーズ）として認識・使用することが知られている．他方で，ニューラル言語モデル (NLM) と人間の言語獲得を対比する一連の研究では，事前に用意された語彙を所与としたものが殆どであり，語彙獲得の段階の研究は十分ではない．そこで本研究では，文字レベルの NLM を用いて，モデルの学習初期においてどのような「かたまり」が形成されていくかを検証した．実験の結果，少なくとも本研究の設定のもとでは，NLM の学習初期にホロフレーズのような現象は観察されなかった．

1　はじめに

人間の乳幼児は，言語獲得の初期段階において，耳にした表現を要素ごとに分解せず，一続きの表現としてそのまま使用することがある [1]．例えば，“I-wanna-do-it” のような発話は，“I”，“wanna”，“do”，“it” のように切り出し可能であるが，言語獲得の初期段階では，そのような分析をすることなく「ひとかたまり」の表現として用いることがある．言語学の用法基盤理論 (usage-based theory) では，このような未分析のかたまりはホロフレーズと呼ばれ，子どもは複雑な文法を理解する前にホロフレーズを認識・使用する段階を経るとされている [2]．

他方で，自然言語処理の分野では，ニューラル言語モデル (NLM) を対象に，言語学や認知科学の観点から言語獲得における人間と NLM の共通点や相違点を検証する研究 [3, 4] が行われており，NLM の学習の効率化などの手掛かりとなる知見が得られてきた．これらの研究の多くは，工学的な動機によって事前に設定された語彙を所与として，その後の言語能力を人間と比較しているが，このような言語能力の基盤となる語の分割や語彙の獲得を考慮しない検証は，特に言語発達の初期段階において，発達的観点から見た妥当性に限界がある．また，NLM の発展以前にも，単語分割能力や語彙の獲得を再現できる計算モデルの探究 [5, 6] が行われてきた．しかし，これらの計算モデルは現代の NLM のように人間に匹敵するような多様な言語的振る舞いを十分に捉えるものではない．

このような背景から，人間と比較可能な水準で様々なタスクにおいて成功を収めた NLM を対象として，言語獲得の初期段階に見られる傾向を明らかにすることが必要である．本研究では，NLM の最初期の発達的特性を理解するための足掛かりとして，言語能力の基盤たる語の分割能力の獲得過程に着目した分析を行う．具体的には，単語単位あるいはサブワード単位の語彙を所与としない文字レベル NLM の学習過程を観察し，先述の用法基盤理論と比較検討することで，NLM が初期にどのような「かたまり」を形成するのかを明らかにする．実験の結果，少なくとも本研究の設定では，ホロフレーズに類する特徴を持つような事例が NLM の学習初期のみで認識される傾向は見られず，NLM の単語分割能力の獲得過程が人間の乳幼児と乖離している可能性が示唆された．

2　実験設定

本研究では，NLM の言語獲得の傾向が人間と比較してどのように異なるか明らかにすることを目的とする．このような目的のもとでは，モデルサイズや訓練データが巨大な最先端のモデルではなく，人間の受け取る入力に近い規模やドメインのコーパスで訓練されたモデルが用いられてきた [7]．このような設定に倣い，人間の学習初期段階に可能な限り設定を揃えて実験を行う．

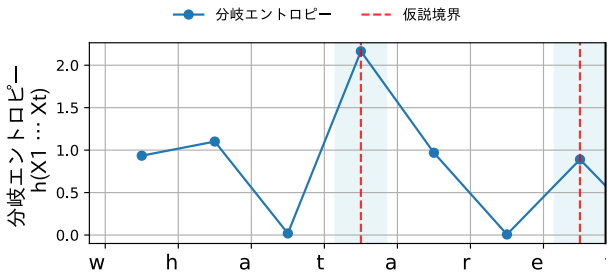

図 1: 系列 $x_0 \cdots x_n$ の各時刻 x_t における分岐エントロピー $h(x_0 \cdots x_t)$ を x_t と x_{t+1} の間にプロットした推移図．赤い点線部分はエントロピーから検出された仮説境界であり，背景の青い部分は真の分節境界である．

2.1 文字レベル NLM の構築

語彙の知識を所与としないために，文字レベルの NLM を構築する．対象とする NLM は，先行研究 [8, 9] の設定に倣い，パラメータサイズ 15M 程度の Llama [8] と GPT-2 [10] とした．詳細な学習設定は付録 B に示す．学習データには，BabyLM Challenge 2023 [11] において、言語獲得段階の幼児が接する言語情報のモデルとして整備された英語コーパスである Baby LM コーパスを用いた．学習データからは単語境界の手がかりとなる空白文字等を削除した．

2.2 NLM による単語境界の推定

本研究では，NLM により計算される確率的不確実性に着目してモデルに内在する単語境界を推定する．確率的不確実性は，幼児も言語獲得で利用している可能性が示唆されている [12]． 本稿では分岐エントロピー (Entropy) [13] に加えて，交差エントロピー損失 (Loss)，予測確率の順位 (Rank)，に基づく不確実性の定式化を行う．

語彙 $\mathcal{V}$ 上の系列 $\boldsymbol{x} = (x_1 \cdots x_{n-1}) \in \mathcal{V}^{|\boldsymbol{x}|}$ と，その次に出現する正解のトークンを $\hat{x}$ としたとき，

Entropy： $h(\boldsymbol{x}) = -\sum_{x \in \mathcal{V}} P_{\mathrm{NLM}}(x|\boldsymbol{x}) \log_2 P_{\mathrm{NLM}}(x|\boldsymbol{x})$

Loss： $h(\boldsymbol{x}) = -\log P_{\mathrm{NLM}}(\hat{x}|\boldsymbol{x})$

Rank： 文脈 $\boldsymbol{x} = (x_1 \cdots x_{n-1}) \in \mathcal{V}^{|\boldsymbol{x}|}$ を受け取ったモデルが語彙全てに対して算出した出現確率のにおける，正解のトークン $\hat{x}$ の確率の順位

と，定義する．また，本研究で使用する境界検出の方法は、以下の 4 つである．

部分系列 $\boldsymbol{x}_n = x_1 \cdots x_n$ と，閾値 $\alpha, \beta \in \mathbb{R}$ に対して，

Relative-Threshold： $h(\boldsymbol{x}_n) - h(\boldsymbol{x}_{n-1}) > \alpha$ のとき，

Absolute-Threshold： $h(\boldsymbol{x}_n) > \alpha$ のとき，

Mixed-Threshold： $h(\boldsymbol{x}_n) - h(\boldsymbol{x}_{n-1}) > \alpha$ かつ，$h(\boldsymbol{x}_n) > \beta$ のとき，

Peak： $h(\boldsymbol{x}_{n-1}) < h(\boldsymbol{x}_n)$ かつ $h(\boldsymbol{x}_n) > h(\boldsymbol{x}_{n+1})$ のとき，

単語境界を検出する．なお，これらの手法は文字レベル LM による単語分割を行う先行研究 [13, 9, 12] における定義を再構成したものである．

2.3 予備実験

本研究の目的は，NLM が文法知識を獲得する過程の初期と人間の乳幼児の言語獲得の様子を比較することであるため，NLM は，妥当な文法知識を獲得している必要がある．そのためまず，NLM の単語分割性能と文法知識の汎化性能を評価する．単語分割性能は Baby LM コーパスの検証データ内から、一部をランダムに抽出したものを用いて評価を行った．正解の単語境界は spaCy[1] によって付与，単語境界推定に用いる閾値 α は開発データから 1%をサンプリングしたものを用いて単語境界推定の F1 値が最も高くなるように調整した．なお，この閾値は後段の分析でも用いた．

NLM による単語境界推定の F1 値は GPT-2 で最大 0.72，Llama で 0.79 であり，本設定における NLM は最終的に十分な単語分割性能を有することを確認した．例として，テキストに対し計算される分岐エントロピーと，推定された分節境界である仮説境界を図 1 に示した．§ 2.2 に記した通り，エントロピーが局所的に増大する箇所を分節境界として推定しており，この事例では NLM によって推定された仮設境界と実際の境界が一致していることが分かる．

文法知識の汎化性能については，BLiMP [14] を用いて評価した．BLiMP は，言語モデルの文法知識を評価するための英語の二値分類ベンチマークである．NLM の BLiMP スコアは 5 エポック目には GPT-2 で 0.66，Llama で 0.67 となり，chance rate の 0.5 を超えて一定程度の文法知識を獲得したことが確認できた．結果の詳細は付録 A に記載している．

3 分析

NLM の学習過程において，モデルの認識する分節境界およびそれらによって形成される分節がどう推移するか，乳幼児のホロフレーズのような分節が現れるかを分析する．

1) https://spacy.io

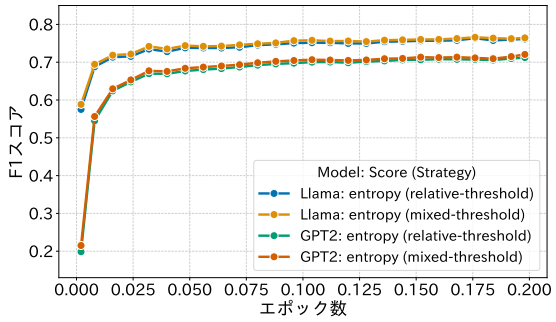

図 2: 最終エポックで F1 が最良であった指標の 0.2epoch までの F1 の推移，非常に初期に分節化能力が獲得され，その後変化しない．

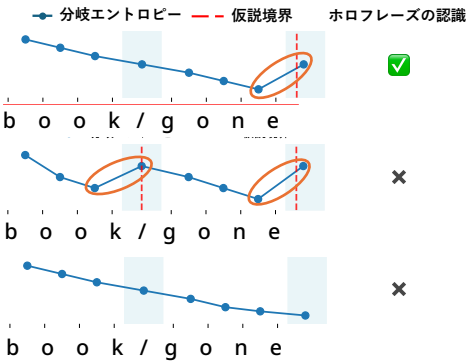

図 3: 分岐エントロピーの推移と NLM がホロフレーズと認識したかの判定. 上段は単語を跨いで境界を認識しているので，ホロフレーズの認識をしていると判定する．中段は境界を単語を跨がず認識しているのでホロフレーズを認識していないと判定する．下段は境界を認識していないのでホロフレーズを認識していないと判定する．

3.1 単語分割能力の推移

まず，学習の各ステップにおいて，§ 2.3 と同様の設定で，単語分割タスクの性能を計測し，NLM がどのように単語分割の能力を獲得するかを観察した．タスクに対する F1 は，GPT-2 で最大 0.72，Llama で 0.79 であり一定程度の分割性能を獲得したことが確認できた．タスクに対する適合率，再現率，F1 値の推移を一部のストラテジ・指標を例に図 2 に示す．図より，学習初期の 1 エポック程度までには急激に分割能力が獲得されていることが分かる．そこで，以降の分析では，単語分割能力が獲得される 1 エポックまでの分節境界の過程を分析する．

この分析で NLM がホロフレーズに類する分割をしているか確認するため，言語学の文献 [15, 16, 17] において乳幼児が使用したことが報告されているホロフレーズの実例を収集し，NLM が認識する分節境界を検出する．ただし，収集できたホロフレーズの数はごく少数であり，定量的な分析を行うには十分でないため，Baby LM コーパスを用いて，分節化の過程を分析する．

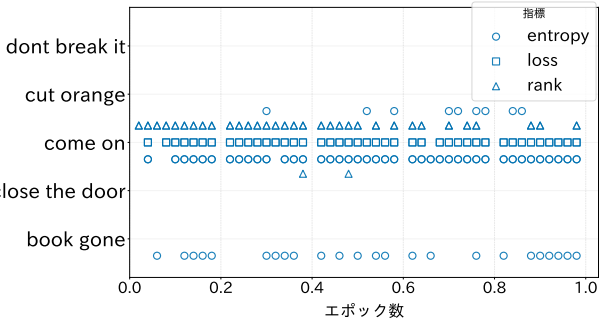

図 4: Llama の mixed-threshold 戦略で検出した各指標毎の結果の一部．ホロフレーズの実例について，NLM が分節化せずホロフレーズと認識した可能性のあるものの，エポック毎の出現位置．人間ように初期に多く出現し，その後分節化されるといった傾向は確認できない．

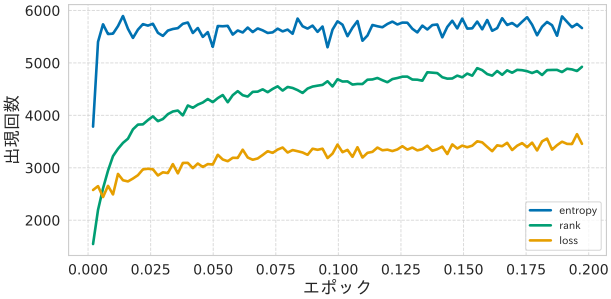

図 5: GPT-2 の mixed-thrshold 戦略における，真の仮説複単語の推移

まず，LM の単語分割能力の獲得過程を観察するために，分節境界検出タスクを学習の各ステップにおいて行なった．タスクに対する F1，精度，検出率は図 2 のとおり．学習初期の 0.1 エポック程度までに急激に分割能力が獲得されることが示された．そこで，次の分析においては，主に最初の 0.2epoch 程度を単語分割能力の獲得段階とみなし，0〜1 エポックを 50 分割して詳細に分析した．

3.2 ホロフレーズの実例を用いた分析

本節では，ホロフレーズの実例に対して NLM がどのように分節境界を見出すかを分析する．そのために，認知言語学の文献 [15, 16, 17] からホロフレーズの実例を収集し，NLM が学習過程でどのように

分節化するか観察した．

ホロフレーズの実例が NLM によって「ひとかたまり」として認識されているか，図 3 に示す基準によって分析する．すなわち，列の終端のみを唯一その系列に存在する分節境界とみなした場合である．分析の結果，ホロフレーズの実例が NLM の学習過程で全ての戦略・指標のうちいずれかによって一度でも「ひとかたまり」として認識された事例は GPT-2 で 5 例，llama で 21 例であった．

ホロフレーズの実例に対して NLM がその学習過程で少なくとも一度以上は「ひとかたまり」であると認識した例について，ほぼ全てのストラテジ・指標の組みで同様の傾向が見られため，Llama で最も F1 が優れていた mixed-threshold 戦略での結果を抜粋して示す．学習過程での出現位置のマップを図 4 に示すとおり，"come on!"のように一貫して「ひとかたまり」と認識されやすいものが見られたものの，複数の指標・戦略にわたって初期のみに見られる事例はほとんど確認できなかった．すなわち，NLM がホロフレーズの実例に対してその学習過程において，乳幼児のように最初期のみ「ひとかたまり」として認識するという特徴は観察されなかった．

3.3 Baby LM コーパスを用いた分析

本節では，収集したホロフレーズの実例ではなく，より一般的なコーパスの文に対してホロフレーズのような性質を持つ事例が NLM の分節として現れるかを検討する．そのために，Baby LM コーパスの開発データから約 1 割の 7,000 文をランダムにサンプリングし，これらの文に対して NLM で推定された仮説境界の分析を行う．

ここで，ホロフレーズの性質は次のように整理でき，本節では，これらの性質を満たすような事例が NLM の学習過程において現れるかを検討する．

性質 1　複数の単語から構成される
性質 2　ひとつの「かたまり」として認識される
性質 3　かたまりとして意味をなす
性質 4　言語獲得の初期段階のみで認識される

まず，与えられたテキストに対し，NLM が検出した仮説境界で囲まれた区間を仮説セグメントと呼ぶこととする．仮説セグメントは NLM が「ひとかたまり」であると認識している系列であると解釈できる．0.2 エポックまでの仮説セグメントの数の推移を図 5 に示す．0.05 エポック目までの初期段階で仮説セグメントの総数が一定の値に収束し，初期に分節が形成されることが確認できる．

これらの分節にホロフレーズに類するものが含まれているのか調べるために，真の単語が複数個連結した仮説セグメントを真の仮説複単語と呼びホロフレーズに類するものとして以降の分析で注目する．

図 5 より，真の仮説複単語の数は，0.05 エポックまでに増加したのちに一定の数を保ちながら推移することが分かる．よって，NLM によってホロフレーズとして認識されている可能性のある事例の数は，学習初期に特に多いわけではなく，人間の乳幼児に見られるような，学習初期のみに認識・使用されるという特徴は観察されなかった．

3.4 議論

分析の結果，少なくとも本研究の設定では，ホロフレーズやそれに類する事例が NLM の学習初期に特に「ひとかたまり」として認識されるという傾向は見られなかった．NLM は学習データの統計的な性質のみに基づいて言語能力を獲得すると考えられるのに対して，人間の乳幼児は社会的相互作用によってホロフレーズを認識・使用すると考えられている [1] ため，この乖離が生じている可能性がある．

4 おわりに

本研究では，NLM の最初期の発達的特性を理解するための足がかりとして，語の分割能力の獲得過程に着目し，文字レベル NLM の学習過程における語の分割の分析を行った．実験の結果，少なくとも本研究の設定では，ホロフレーズに類する事例が NLM の学習初期に「ひとかたまり」として認識される一貫した傾向はみられず，NLM の単語分割能力の獲得過程が人間の乳幼児とは乖離している可能性が示唆された．

一方で，本研究ではこれらの現象について初期に確認できるかという定性的な評価のみ行った．定量化による一般性を担保は今後の課題であるまた，本研究で主に注目した仮説セグメントは真の仮説複単語のみであるが，単語以外にも意味のかたまりとして妥当な単位はあり得る．他の仮説単語についても，意味のある接頭辞や接尾辞の単位で境界が推定されている可能性がある．これらの分析も今後の課題とする．

参考文献

[1] Michael Tomasello. **Constructing a Language: A Usage-Based Theory of Language Acquisition**. Harvard University Press, 2003.

[2] Michael Tomasello. Do young children have adult syntactic competence? **Cognition**, Vol. 74, No. 3, pp. 209–253, 2000.

[3] Ionut Constantinescu, Tiago Pimentel, Ryan Cotterell, and Alex Warstadt. Investigating critical period effects in language acquisition through neural language models. **Transactions of the Association for Computational Linguistics**, Vol. 13, pp. 96–120, 2025.

[4] Linnea Evanson, Yair Lakretz, and Jean-Rémi King. Language acquisition: do children and language models follow similar learning stages? In **Findings of the Association for Computational Linguistics**, pp. 12205–12218, Toronto, Canada, 2023.

[5] Mark Johnson Sharon Goldwater, Thomas L. Griffiths. A bayesian framework for word segmentation: Exploring the effects of context. **Cognition**, Vol. 112, No. 1, pp. 21–54, 2009.

[6] Padraic Monagahan and Morten H. Christiansen. Words in puddles of sound: modelling psycholinguistic effects in speech segmentation. **Journal of Child Language**, Vol. 37, No. 3, p. 545–564, 2010.

[7] Michael Y. Hu, Aaron Mueller, Candace Ross, et al. Findings of the second BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In Michael Y. Hu, Aaron Mueller, et al., editors, **The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning**, pp. 1–21, Miami, FL, USA, November 2024. Association for Computational Linguistics.

[8] Hugo Touvron, Thibaut Lavril, Gautier Izacard, et al. Llama: Open and efficient foundation language models. **arXiv**, 2023.

[9] Zebulon Goriely and Paula Buttery. BabyLM's first words: Word segmentation as a phonological probing task. In Gemma Boleda and Michael Roth, editors, **Proceedings of the 29th Conference on Computational Natural Language Learning**, pp. 522–539, Vienna, Austria, July 2025. Association for Computational Linguistics.

[10] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. **OpenAI blog**, 2019.

[11] Alex Warstadt, Leshem Choshen, Aaron Mueller, Adina Williams, Ethan Wilcox, and Chengxu Zhuang. Call for papers – the babylm challenge: Sample-efficient pretraining on a developmentally plausible corpus, 2023.

[12] Yugo Murawaki. Unsupervised synthesis of word language models from pretrained character language models. **IPSJ SIG Technical Report**, Vol. 2024-NL-260, No. 2, pp. 1–14, June 2024.

[13] Kumiko Tanaka-Ishii and Zhihui Jin. From phoneme to morpheme: another verification using a corpus. In **Proceedings of the 21st International Conference on Computer Processing of Oriental Languages: Beyond the Orient: The Research Challenges Ahead**, ICCPOL'06, p. 234–244, Berlin, Heidelberg, 2006. Springer-Verlag.

[14] Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. BLiMP: The benchmark of linguistic minimal pairs for english. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 377–392, 2020.

[15] Elena V. M. Lieven, Julian M. Pine, and Helen Dresner Barnes. Individual differences in early vocabulary development: redefining the referential-expressive distinction. **Journal of Child Language**, Vol. 19, No. 2, p. 287–310, 1992.

[16] Michael Tomasello. **Constructions**, Vol. Special Volume 1, pp. 1–23, 2006.

[17] Janet L. Patterson. Development of constructed phrases in a child with language impairment. **Clinical Linguistics & Phonetics**, Vol. 14, No. 7, pp. 545–556, 2000.

[18] Julian M. Pine and Elena V. M. Lieven. Reanalysing rote-learned phrases: individual differences in the transition to multi-word speech. **Journal of Child Language**, Vol. 20, No. 3, p. 551–571, 1993.

A　BLiMP スコア

BLiMP スコアの詳細を，図 6 に示す．図表に示した精度は 67 タスクスコアごとの精度をさらに平均したものである．

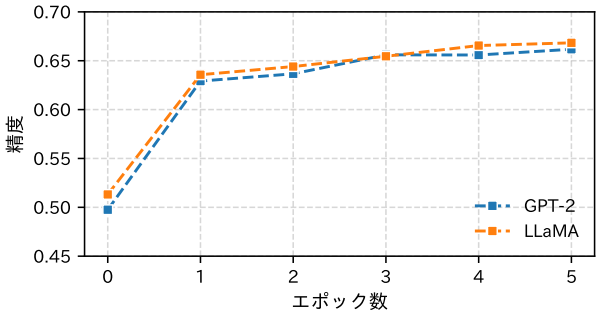

図 6: 各エポックにおける BLiMP 正解率 (67 タスク正解率の平均)．GPT-2，Llama のどちらも，学習のごく初期段階でスコアが上がりきっている．本研究では初期段階のみに注目していたが，それは初期段階であっても十分文法知識を備えていることが確認できたことからも妥当な設定と言える

B　実験設定

本研究で作成した言語モデルのハイパーパラメータを，表 1 に示す．

分析で用いたホロフレーズの実例を表 2 に示す．これらの実例は，Tomasello [16] の holophrase の例の他に，Pine ら [18] の frozen phrase の例，Patterson [17] の routine phrase の例から取得した．

Tomasello [1] は “I-wanna-do-it” のような例を挙げ，子どもがそのような表現を要素に分解せず，holophrase として学習を開始すると述べている．そのことの実証的な裏付けのために，Pine ら [18] による frozen phrase に関する研究を参照しており，holophrase と frozen phrase は同一の概念であることが伺える．また，Patterson [17] による研究では，routine phrase も frozen phrase と同列に扱われていることから，本研究の分析対象に含めた．

なお，Tomasello [1] の言及する holophrase には，“I-wanna-do-it” のように複数の単語から構成されるものだけでなく，“towel” のように単一の単語から構成されるものも含まれる．したがって本研究では，複数の単語から構成される holophrase や frozen phrase，routine phrase を対象として収集した．

表 1: ハイパーパラメータ

		Llama	GPT-2
Model	architecture	Llama	GPT-2
	parameters	15M	15M
	vocab size	32	32
	context size	512	512
	hidden size	512	512
	n-inner	-	768
	heads	8	8
	layers	8	8
	layer norm eps	1e-05	1e-05
Optimizer	learning rates	3e-4	3e-4
	weight decay	0.01	0.01
Training	gradient accumulation step	4	4
	epoch	5	5
	batch size	16	16

表 2: 本研究で収集したホロフレーズの実例

実例	文献	文献中の用語
Oh dear	[18]	frozen phrase
Oh God !	[18]	frozen phrase
Where’s it gone	[18]	frozen phrase
There it is	[18]	frozen phrase
Book gone	[18]	frozen phrase
Daddy gone	[18]	frozen phrase
Pick up !	[16]	holophrase
Lemme see !	[16]	holophrase
I wanna do it	[16]	holophrase
lemme have it	[17]	routine/frozen phrase
open the door	[17]	routine phrase
push it right there	[17]	routine phrase
lemme have it	[17]	routine phrase
uh need a knife	[17]	routine phrase
give to me	[17]	routine phrase
cut orange	[17]	routine phrase
you OK	[17]	routine phrase
need a knife	[17]	routine phrase
close the door	[17]	routine phrase
I turn	[17]	routine phrase
right there	[17]	routine phrase
I don’t know	[17]	routine phrase
come on	[17]	routine phrase
uh go try it	[17]	routine phrase
try it	[17]	routine phrase
door open	[17]	routine phrase
two juice	[17]	routine phrase
give it to me	[17]	routine phrase
don’t break it	[17]	routine phrase
here is	[17]	routine phrase
move your fingers	[17]	routine phrase
come on	[17]	routine phrase