

ロボットによるオンライン語彙学習 —ロボットが一週間で学んだものとは?—

Online Learning of Words by Robots: Results of Word Learning by a Robot in One Week

荒木 孝弥 中村 友昭 長井 隆行
Takaya Araki Tomoaki Nakamura Takayuki Nagai

電気通信大学大学院情報理工学研究科

Faculty of Informatics and Engineering, The University of Electro-Communications

One of the biggest challenges in intelligent robotics is to build robots that can understand and use language. Such robots will be a part of our everyday life; at the same time, they can be of great help to investigate the complex mechanism of language acquisition by infants in constructive approach. To this end, we think that the practical long-term on-line concept/word learning algorithm for robots and interactive learning framework are the key issues to be addressed. For the on-line learning algorithm we develop a practical method that solves three remaining problems in our previous study. We also propose an interactive learning framework, in which the proposed on-line learning algorithm is embedded. The main contribution of this paper is to develop such a practical learning framework, and we test it on a real robot platform to show its potential toward the ultimate goal.

1. はじめに

ロボットによる概念・言語の獲得や理解は、知能ロボティクスの大きな課題の一つであろう [Iwahashi 07]. このための重要な要素は、経験の分類と、その分類を通じた予測である。従来、人間の認知機能においてもカテゴリ分類の重要性が指摘されている。事実、人間は様々な情報をその類似性によってクラスタリングすることでカテゴリ (概念) を形成している。また、人間が用いている単語はカテゴリに基づいており、ロボットもカテゴリ分類を通じて物体の概念を学習することで、未観測情報の予測や言語の理解が可能になると考えられる。

我々は、ロボットが取得したマルチモーダル情報を自律的に分類し、人間の発話した単語をカテゴリと結びつけることで、物体の概念や語意が獲得できると考え、統計的教師なしクラスタリング手法であるマルチモーダル LDA を提案した [Nakamura 09]. そしてそのオンライン化や、教師なし形態素解析器との統合などを進めてきた [Araki 12]. 本稿では、この枠組みを実ロボットに搭載し、実際に多数の物体を学習させた実験について述べる。その結果、実ロボットが音素知識のみからどの程度学習できたのかについて考察する。

本稿ではまず、[Araki 12] のアルゴリズムを概観するとともに、問題点を挙げそれらを改善する手法を提案する。具体的には、機能語が多く予測される問題への対処、言語モデル学習の疑似オンライン化、音素認識誤りに対する対処の3つである。また、後半ではロボット-教示者間のインタラクション設計とその実装について述べる。学習者と教示者との間のインタラクションの重要性は、従来から認識されている。例えば [Roy 09] は、実生活における子供と養育者のインタラクションを長期間に渡り解析し、その根底にあるダイナミクスを明らかにした。一方、ロボットのアフォーダンス学習という文脈において、教示者の行動がロボットの学習に影響を与えること、また逆にロボットの学習が教示者の行動に影響を及ぼすことが示されている [Thomaz 09]. 本稿でも、ロボットの語彙学習にインタラクションの枠組みを適用することで、ロボットの学習がどのように変化するかを検証する。実験として、インタラクティブな学習を観察する実験と、一週間でロボットがどの程度の語彙を学習できるかを検証する実験の二つを行い、その結果を通して、

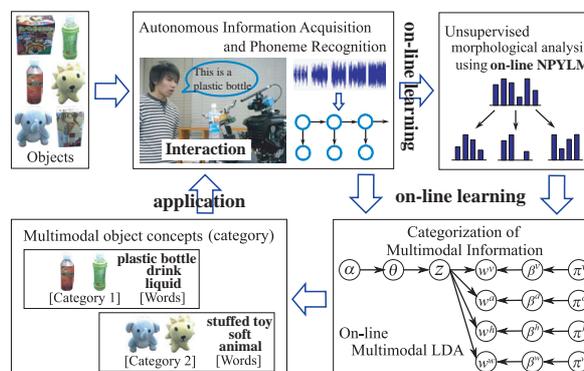


図 1: システム概要図

ロボットの長期的な言語学習の可能性を示す。

2. オンライン語彙学習

2.1 システムの概要

オンライン学習システムの概要を図 1 に示す。学習アルゴリズムは、図 1 中に示すグラフィカルモデルに基づいており、物体のカテゴリゼーションはそのパラメータを推定する問題として考える。本稿では、入力データとして、ロボットが自律的に取得可能な視覚、聴覚、触覚情報と、ユーザーとのインタラクションによって得られる単語情報を用いる。以下に具体的な概念形成手法について述べる。

2.2 マルチモーダル情報取得システム

本稿で想定するロボットを、図 2 (a) に示す。ロボットは 6 自由度のアーム、CCD カメラ、赤外線カメラ、マイク、触覚センサーを搭載しており、自律的に情報取得を行うことができる。視覚情報として、物体を複数視点から観測することで得られる画像情報 (図 2 (b)), 聴覚情報として物体を振った際に得られる音情報 (図 2 (c)), 触覚情報として物体を把持した際に得られる感圧情報 (図 2 (d), (e)) を取得することが可能である。また、各情報を取得している最中に、人が物体に関する発話をするので、音声認識により教示発話を取得することが可能である。これらの情報は全て Bag of Features (BoF) モデルとして扱い、位相に依存しない生起回数の情報としてカテゴ

連絡先: 長井 隆行, 電気通信大学大学院情報理工学研究科, 東京都調布市調布ヶ丘 1-5-1, taraki@apple.ee.uec.ac.jp

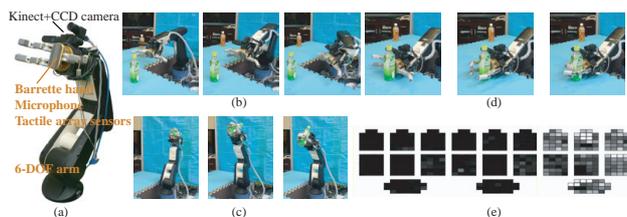


図 2: ロボットプラットフォームと自律的情報取得

り分類に利用する。視覚情報には、物体を複数の視点から観測した際に得られる 128 次元の DSIFT 特徴量を用い、各画像から計算された特徴ベクトルをベクトル量子化することで得られる 500 次元のヒストグラムを用いる。聴覚情報には、物体を振った際に得られる音情報から MFCC を計算し、これをベクトル量子化することで得られる 50 次元のヒストグラムを、触覚情報には、ハンドに搭載された 162 個の感圧センサから取得される時系列データの近似パラメータをベクトル量子化して得られる 15 次元のヒストグラムを用いる。

2.3 オンラインマルチモーダル LDA

本稿で提案するモデルは、図 1 中のグラフィカルモデルに相当する。図 1 において、各情報 w^v , w^a , w^h , w^w は、それぞれ視覚、聴覚、触覚、単語情報を表し、これらの情報は先に述べた情報取得システムと後述の Nested Pitman-Yor 言語モデル (NPYLM) により取得される。提案モデルにおいて、これらのマルチモーダル情報 w^* は、ハイパーパラメータ π^* によって決まるディリクレ事前分布に従うパラメータ β^* の多項分布から発生する。また、 z はカテゴリを示し、ハイパーパラメータ α によって決まるディリクレ事前分布に従うパラメータ θ の多項分布により生成される。カテゴリ分類は、取得した知覚情報に基づきモデルのパラメータ θ と β を推定することに相当し、パラメータ推定にはギブスサンプリングを適用する。[Araki 12] では、これをさらにオンライン化し、ロボットによる逐次的な学習を可能とした。基本的な考え方は、忘却パラメータを設けることでモデルをアップデートしていくことであるが、さらにパラメータを複数用意し、この中から単語の予測性能が高いと思われるものをパーティクルフィルタによって選択する。学習の詳細については、[Araki 12] を参照されたい。

モデルに基づいた未観測情報の予測は、ベイズ推定によって実現することができる。例えば、物体に関する知覚情報から単語を予測するためには、

$$P(w^w | w^*) = \int \sum_z P(w^w | z) P(z | \theta) P(\theta | w^*) d\theta \quad (1)$$

を計算し、確率の大きなものを選択すればよい。ただし、単語の予測に関しては、この計算を単純に行うと、「です」や「これ」といった、機能語が常に高い確率で予測されることになる。それを避けるために、TF-IDF (Termed Frequency - Inverse Document Frequency) を利用する。

2.4 階層ベイズ言語モデル

カテゴリ分類で利用する単語情報は、前節で示した情報取得システムにより取得された教示発話に対して、形態素解析を行うことで取得可能である。この場合、事前にロボットが単語辞書を保持していると仮定し、教示発話を形態素解析器によって単語へと分割し、Bag of Words (BoW) 表現とすることでカテゴリ分類に利用することとなる。しかし、形態素解析器を用いる場合、辞書に登録されていない未知語への対応ができず、ロボットには単語の切り出しを自律的に行う枠組みが必要であると言える。そこで [Araki 12] では、[Mochihashi 09]

で提案された階層ベイズ言語モデルである NPYLM を導入することで、教師なしで形態素解析を行うことを提案した。

NPYLM は、単語 N -gram モデルと文字 N -gram モデルを併用し、各 N -gram モデルに Pitman-Yor 過程によるスムージングを行うことで、未知語や低頻度語に対するロバスト性を向上させる手法である。また、この言語モデルを利用しブロック化ギブスサンプリングと動的計画法を用いて高速に単語の分割を行い、階層 Pitman-Yor 過程を用いることで、入力データのみから教師なしで音素列の分節化が行える。この手法により、ロボットは事前知識を用いず、音素モデルのみによって音素認識を行い、得られた音素列から単語を教師なしで切り出すことが可能となる。最終的に、ロボットは音素認識によってユーザーの教示発話を認識し、NPYLM による単語の切り出しを行うことで得られる単語を BoW 表現とすることでカテゴリ分類に利用する。

ここで問題となるのは、NPYLM がバッチタイプのアルゴリズムであることである。[Araki 12] では、新たな発話が入力されるたびに過去の全ての音素列とともに NPYLM を実行することで言語モデルの更新を行っている。しかし保持している音素列が多くなると計算に時間がかかるため、教示された発話の情報をすぐに生かすことができず、インタラクティブな学習では問題となる。そこでここでは、NPYLM を疑似的にオンライン化することを考える。アイデアは非常に単純であり、NPYLM の計算に用いる発話数を一定に保ち (後に示す実験では 100 発話とした)、それ以前の発話は計算に用いないことで計算量が増えることを防ぐというものである。ただし前回の学習結果を学習時の初期値とする。

もう一つの重要な問題は、音声認識の際の音素誤りである。[Araki 12] においても音素誤りについて検討しており、音素誤りが 20% 程度であれば分類に大きな影響がないことを示している。しかし、実際の学習では音素誤りが 20% を超えることが多く、このことが学習の大きな妨げとなる。そこで本稿では、音素誤りを含む音素列を NPYLM により単語分割し、BoW を生成する際に編集距離を考慮した投票を行うことでこの問題に対処する。投票するための重みは、

$$\text{voting weight} = \frac{(\text{word length} - \text{edit distance})}{\text{word length}}, \quad (2)$$

として計算する。これにより、同一単語の一部の音素が誤ることによって異なる単語として投票されてしまうことを防ぎ、同時に音素が大きく誤った単語は、一度のみ投票されることでその割合が小さくなる。ただし、音素誤りによって生じた単語も全て辞書に登録されるため、入力情報から単語を予測する際には、多くの似た単語が予測されることになる。そこで、単語を予測した結果に対して、上記の重み付き投票を行ない、確率の最も高い単語を選択し、その単語に編集距離の近い単語を全て破棄する、といった手順を繰り返すことで予測単語を選択する。

2.5 インタラクティブ学習のフレームワーク

概念・言語学習において、他者とのコミュニケーションは非常に重要な要因であり、本研究で想定するシナリオにおいても、ユーザーとのインタラクションは必要不可欠である。本稿では、語彙学習におけるインタラクティブな学習フレームワークを提案する。実際に構築した学習フレームワークの概要を、図 3 に示す。ロボットの行動はメイン行動部 (図 3 左) と予測行動部 (図 3 右) によって構成される。メイン行動では、まずロボットが人物の探索を行い、ユーザーがいる場合にはインタラクション行動を取る。本稿では、インタラクション行動として幼児の行動を模倣した共有注意行動や視線の検出、対話に対する予測行動などを想定した。一方、ユーザーが発見できな

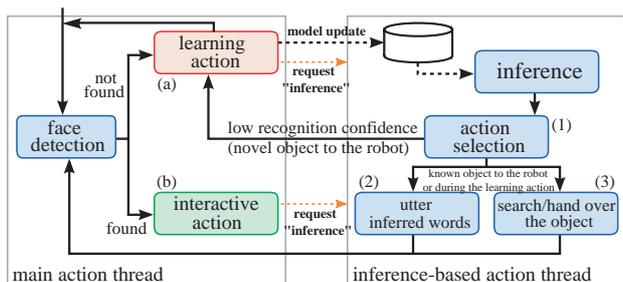


図 3: 実環境インタラクティブオンライン学習フレームワークのアーキテクチャ



図 4: 実験に使用した 125 個 24 カテゴリの物体

い場合には、提案する情報取得システムとオンライン学習システムにより、ロボットは自律的な学習を行う。また、ロボットはメイン行動と並行して、図 3 右に示した様々な予測行動を取る。ここでは予測行動として、知覚情報に基づく単語の予測と発話、単語情報に基づく視覚情報の予測と物体の探索などを想定した。本稿で提案するフレームワークの実装により、ロボットは実環境でユーザーとのインタラクションを行いながらオンラインで概念と語彙の更新を行うことが可能となる。

3. 実験

提案する学習アルゴリズムを図 2 のロボットに実装し、インタラクティブ学習の観察と一週間の語彙学習実験を行った。実験には、図 4 に示す 125 個 24 カテゴリの物体を用いた。分類精度は人手による分類を正解とし、正解との差が最も小さくなるように、分類したカテゴリ ID を並べ替えた際の誤差を評価した。図 5 に実験環境とインタラクションの様子を示す。

3.1 学習におけるインタラクションの観察

4 人の被験者にロボットに単語を教えるように指示をし、実験を行うために必要となる簡単な説明を行った。4 人はいずれも、本研究のロボットについては事前の知識を持っていない 20 代男性であり、うち 2 名は提案したインタラクティブな学習フレームワークで、残りの 2 名はインタラクションなしの条件で学習実験を行った。この実験では、図 4 の物体からランダムに 20 個 5 カテゴリの物体を選び使用した。インタラクションなし条件では、被験者は 20 個の物体を順に一つずつロボットに教えるが、ロボットは発話などのインタラクションを一切行わない。インタラクションあり条件では、被験者はどの物体をどの順番で何回教えるかを自由に決めることができることとした。ただし 20 回教示を行った段階で、全ての物体を教え終わっていても実験を終了とした。この実験におけるすべての被験者発話にはラベル付をし、平均発話長 (Mean

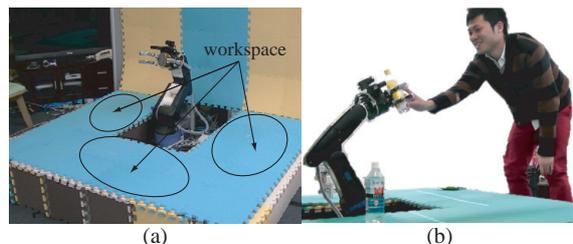


図 5: インタラクティブ学習実験の様子, (a) ロボットとワークスペース, (b) インタラクティブ学習の様子

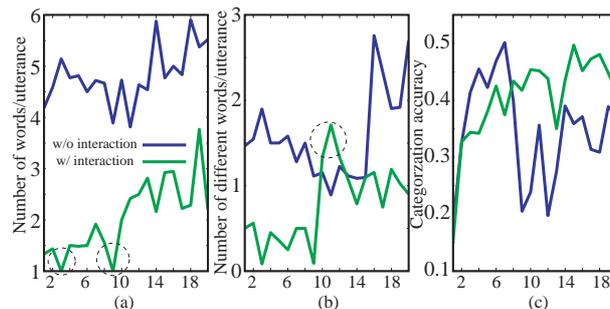


図 6: インタラクティブ学習実験の結果: (a) MLU, (b) TTR, (c) カテゴリ分類性能

Length of Utterances: MLU) とタイプ・トークン比 (Type Token Ratio:TTR) を計算した。図 6 に結果を示す。ただし、各条件の 2 名は同様の傾向を示したため、一名ずつの結果を示している。

MLU は文が平均何語から構成されているかを意味しており、図 6 (a) よりインタラクション条件では、徐々に文の複雑さが増していることが見て取れる。これは、ロボットとのインタラクションにより、どの程度学習が進んでいるかを把握できるため、その進度に合わせて複雑さが調整されている可能性を示している。実際、図 6 (c) のカテゴリ分類精度と MLU の間にはインタラクション条件の時のみ有意な正の相関が確認された ($r = .50, p < .05$)。また、図 6 (a) において丸で示した MLU の落ち込みは、次のように解釈できる。例えば、9 回目での大きな落ち込みは、9 回目の教示でロボットが正しく物体名を発話しており、その直後から教示者の発話は複雑化していった。9 回目の直前では、ロボットはなかなか正しく対象としての物体に関する発話をしなかったため、教示者は発話の複雑さを減らし、結果として一語発話に近づいている。また、図 6 (b) の TTR でも 10 回目の教示で大幅に値が増加していることが分かる。これは、ロボットがあることを正しく覚えたことが確認できたため、より多くのことを複雑なやり方で教えることができる、という教示者の意図が関与していると思われる。こうした、学習者と教示者の間のダイナミクスは、時間スケールの違いこそあれ、[Roy 09] で明らかにしている知見とも一致しており興味深い。

3.2 オンライン語彙学習実験

3.2.1 ロボットの獲得した語彙

図 4 の全ての物体を使ったオンライン語彙学習実験を行った。教示者は一人であり、合計 200 回の教示を 1 日 3~5 時間で 1 週間かけて行った。この実験において、教示者は合計 1055 発話行い、ロボットは 924 単語を獲得 (辞書に追加) した。ただし、これには音素誤りや切り出しの誤りによる意味不明な単語が 632 単語含まれていた。意味の分かる単語は、計 58 単語

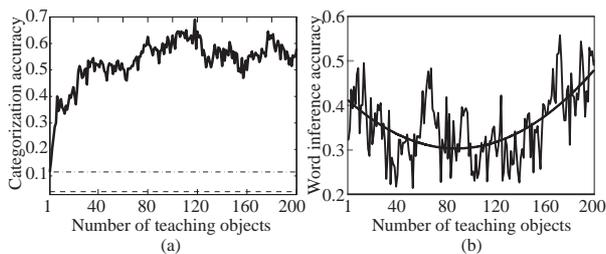


図 7: 学習回数に応じたカテゴリ分類精度と単語予測精度

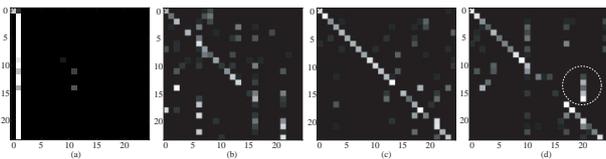


図 8: 学習回数に応じたカテゴリ分類結果 (混同行列): (a) 1 回目の学習結果, (b)10 回目の学習結果, (c) 120 回目の学習結果, (d) 200 回目の学習結果

あり, そのうち 4 単語は機能語, 10 単語は形容詞, 40 単語は名詞, 4 単語は動詞であった. 残りの 234 単語は, 58 単語に近く音素列の一部が異なる重複で, これらは編集距離によって省くことができるものである. また, 提案した編集距離による重み付き投票により, 無意味な単語への得票数を抑えることができ, BoW におけるその割合は 10%程度であった.

3.2.2 分類精度の評価

図 7 (a) に各学習回数に対する分類精度の結果を示す. これより, 学習とともに段階的に分類精度が向上していることが分かる. しかし, 120 回程度の教示後は向上が見られず, 緩やかに振動していることが分かる. これは, 知覚情報の似通ったいくつかのカテゴリの存在が原因であると思われる. 例えば, “スナック菓子”と“クッキー”はいずれもパッケージの箱であり, 見た目や触覚での判別は難しい. こうしたいくつかの分類が難しいカテゴリが合併, 分離を繰り返すことで全体の精度が振動している. 図 8 (d) の丸で示したカテゴリは, そうした分類の難しいカテゴリであり, 200 回目では合併しているが, 120 回目ではうまく分類できていることが分かる. こうした分類困難なカテゴリが, 学習を継続することで分類可能となるかどうかは, 今後の継続実験の結果を待つ必要がある.

3.2.3 未知物体に対する単語予測

提案手法による未知物体に対する単語の予測実験を行った. テストセットとして, 学習に用いていない各カテゴリ 1 物体の計 24 物体を用意した. ロボットが各物体を見ることで単語を予測し, その単語が物体を表現するのにふさわしいかどうかを人手で判断し, ふさわしいと判断された割合を精度とした. 実験結果を図 7 (b) に示す.

結果は概ね U 字形を示しているが, これは次のように解釈できる. 学習初期は, 教示された単語が少なく, 機能語が多くのカテゴリに結びついている. この段階では, TF-IDF が十分に機能せず, 予測される単語の多くは機能語となっている. 機能語は表現として間違いではないため, 予測精度としては比較的高い値となる. 一方, 学習が進み単語が多くなると予測精度は下がるが, 学習がさらに進むことで予測精度が徐々に向上する.

図 9 は学習に用いていないコップを見せた際にロボットが予測した単語の例である. この例では, 正しく“コップ”という単語が最も高い確率で予測されている. しかし, TF-IDF

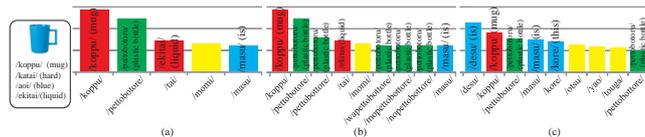


図 9: 概念モデルを用いて予測された単語の例: (a) 提案手法, (b) 編集距離を考慮しない場合, (c)TF-IDF を用いない場合 (赤: 正解, 緑: 不正解, 黄色: 意味不明語, 青: 機能語)

を用いていない (c) では, “です”が最も高い確率で予測されている. また, 編集距離を用いていない (b) では, “ペットボトル”に付随した多くの重複語が予測されている. “ペットボトル”自体は誤った予測であり, 提案手法でも 2 番目に高く予測されているが, 編集距離を考慮して結果をマージするため, 似通った多くの単語を予測してしまうことがない.

実験では, どのように教示発話を行うかは一切規定していないため, 当然物体の色に関する教示も含まれている. しかし, 視覚情報として色の情報を用いていないため, 原理的に色の概念を正しく形成することができない. あるカテゴリに偶然似た色の物体が集まった場合に, そのカテゴリに形容詞を結びつけてしまうことがあり, その結果, 入力された物体とは全く関係のない色を発話するケースが見られた. この問題はどのような特徴量を用いるかに依存し, 色の場合は色の情報を視覚特徴に付加しなければ解決することができない. 基本的には, 考え得るすべての特徴量を抽出し, どの特徴量を使って分類するかを自動的に判断するような仕組みが必要である. [Nakamura 12] ではこの問題について検討しており, 今後こうしたアイデアをオンライン学習に取り入れる必要がある.

4. 結論

本稿では, ロボットによる物体概念・言語獲得のためのオンライン学習アルゴリズムを提案した. この手法をインタラクティブ学習のフレームワークにより統合し, 実機ロボットを用いた実験によって長期間の言語学習の可能性を示した. 現在さらに実験を継続しており, その結果を解析することや, 被験者数を増やしたインタラクション観察実験が今後の課題である.

参考文献

[Araki 12] Araki, T., *et al.*: “Online Learning of Concepts and Words Using Multimodal LDA and Hierarchical Pitman-Yor Language Model”, in Proc. of IROS, pp.1623–1630, 2012

[Iwahashi 07] Iwahashi, N.: “Robots That Learn Language: A Developmental Approach to Situated Human-Robot Conversations”, in N.Sankar ed. Human-Robot Interaction, pp.95–118, I-Tech Education and Publishing, 2007

[Mochihashi 09] Mochihashi, D., *et al.*: “Bayesian Unsupervised Word Segmentation with Nested Pitman-Yor Language Modeling”, in Proc. of ACL-IJCNLP, volume. 1, pp. 100–108, 2009

[Nakamura 09] Nakamura, T., *et al.*: “Grounding of Word Meanings in Multimodal Concepts Using LDA”, in Proc. of IROS, pp. 3943–3948, 2009

[Nakamura 12] Nakamura, T., *et al.*: “Bag of Multimodal Hierarchical Dirichlet Processes: Model of Complex Conceptual Structure for Intelligent Robots”, in Proc. of IROS, pp.3818–3823, 2012

[Roy 09] Roy, B. C., *et al.*: “Exploring Word Learning in a High-Density Longitudinal Corpus”, in Proc. of the 31st Annual Meeting of the Cognitive Science Society, pp.2106–2111, 2009

[Thomaz 09] Thomaz, A. L., *et al.*: “Learning about Objects with Human Teachers”, in Proc. of HRI’09, pp.15–22, 2009