# Discovering the typing behaviour of Parkinson's patients using topic models

Antony Milne[1], Mihalis Nicolaou[1,2], and Katayoun Farrahi[1]

[1] Goldsmiths, University of London, UK
[2] Imperial College London, UK

**Abstract.** Sensing health-related behaviours in an unobtrusive, ubiquitous and cost-effective manner carries significant benefits to healthcare and patient management. In this paper, we focus on detecting typing behaviour that is characteristic of patients suffering from Parkinson's disease. We consider typing data obtained from subjects with and without Parkinson's, and we present a framework based on topic models that determines the differing behaviours between these two groups based on the key hold time. By learning a topic model on each group separately and measuring the dissimilarity between topic distributions, we are able to identify particular topics that emerge in Parkinson's patients and have low probability for the control group, demonstrating a clear shift in terms of key stroke duration. Our results further support the utilisation of key stroke logs for the early onset detection of Parkinson's disease, while the method presented is straightforwardly generalisable to similar applications.

**Keywords:** health behaviour models, topic models, Latent Dirichlet Allocation

## 1 Introduction

Early diagnosis of progressive neurodegenerative disease plays a crucial role in maximising the impact of medication and preventing (as far as possible) further progression of the disease. In particular, Parkinson's disease is considered a slow progressing neurodegenerative disease. While diagnosis is usually performed by considering the patient's symptoms as well as a physical examination, a main characteristic of Parkinson's disease lies in the manifestation of motor symptoms during the early stage of the disease. Since the cost of constantly monitoring motor signs can be prohibitive for healthcare systems, a possible alternative is to study the daily behaviours of patients. Typing behaviour can be indicative of degenerative motor signs during the early stages of the disease. This constitutes an unobtrusive, ubiquitous, transparent and inexpensive approach, since data can be collected while subjects perform their daily routines.

Motivated by the above, in this paper we study the typing behaviour of Parkinson's patients in an unsupervised setting, and contrast results to a control group. To this end, we consider topic models, and in particular Latent Dirichlet

Allocation (LDA) [2], a statistical generative model that has been successfully employed in human behaviour mining problems in various contexts [4, 7–9, 11]. LDA discovers semantically coherent latent *topics* in a collection of data (called a *corpus*), with low-level features (or *words*) being generated from a distribution of topics. Considering a recently released dataset of typing data [5], we present an approach where LDA models are trained on each of the Parkinson's and control groups. We consider a typing session to be analogous to a *document*, with the key hold durations corresponding to words. We show that discovered topics are heavily shifted towards long key hold times in the Parkinson's patient group, and that discovering topics containing long key press durations is likely to be an indicator of the disease. While key hold times are the dominant feature considered in this paper, the methodology proposed can easily extend to words encoding multiple features. We further validate our findings by computing the similarity between topics across groups with the Bhattacharyya coefficient. The results of this analysis indicate which topics, and therefore which words and co-occurrence of words, have the highest probability of being associated with Parkinson's disease.

## 2 Related work

Machine learning techniques have previously been used to monitor and automatically detect the severity of Parkinson's symptoms considering speech data [15, 6]. Home monitoring systems for Parkinson's patients have also been developed using accelerometer data [3] as well as gyroscope data sensing upper body activities [13]. Wearable sensor data has been used to estimate the severity of Parkinson's symptoms, such as tremor, bradykinesia and dyskinesia from accelerometer data features [12]. While most works in the wearable sensing community consider Parkinson's monitoring using upper body sensors, shoe worn sensors have also been used to assess locomotion for early diagnosis [10]. Bachlin et al. [1] use wearable technology to study gait, particularly freezing of gait by using accelerometer sensors attached to the belt and lower body.

Giancardo et al. [12] recently proposed to use typing data for the detection of Parkinson's disease. This is the first study to consider typing data and it has the major advantage of "transparency" and ubiquitousness. Giancardo et al. propose a neuroQWERTY index (nQi) based on key hold times in order to classify Parkinson's vs. control patients. In this work, we consider a very different approach to the problem, one based on unsupervised learning, and the goal is to discover the precise patterns which may be discriminative of Parkinson's disease. Our proposed approach is generalisable to new features that can be obtained by typing and to other similar problems.

## 3 Methodology

All the analysis performed in this paper uses the datasets MIT-CS1PD and MIT-CS2PD associated with Ref. [5]. Subjects in Madrid, Spain performed typing tests

by transcribing a folk tale on a word processor for 15 minutes with key stroke data being recorded. Some subjects completed two such typing tests. For our purposes, the key strokes from each individual typing test form a *document*. In total there are 116 documents, of which $D^P = 60$ correspond to Parkinson's sufferers and $D^C = 56$ to control subjects. These two collections of documents correspond to our groups of Parkinson's and control. Using the language of LDA [2], each collection of documents forms a *corpus*.

The documents are preprocessed to keep only key presses corresponding to alphabetic and the space, comma, period and return keys. Thus any extraneous key presses such as shift, accents and backspaces are removed. To filter out erroneously recorded data (e.g., arising from simultaneous or overlapping key strokes) we keep only key strokes whose duration is between 0 and 0.5 seconds, and whose travel time to the following key is between 0.02 and 20 seconds. Following this data cleaning procedure, the average number of key strokes in a document is 1240 (standard deviation 470).

We next convert each document into a collection of *words* that will be analysed by the topic model. Again, we emphasise that a 'word' in this context does not describe the actual Spanish word typed, but instead relates to the duration for which each key is pressed. Combining the control and Parkinson's corpora, we find a set of boundaries that will allocate each key press duration into one of 50 bins. These boundaries are calculated so that there is a roughly equal number of key strokes in each bin across the two corpora. Owing to the distribution of key press durations, the bin widths are not uniform. For example, bin number 5 corresponds to a key press between 0.0615 and 0.0645 seconds, whilst bin number 45 corresponds to a key press between 0.1697 and 0.1771 seconds. The word that labels each key stroke is given by the number of the key press duration bin to which it is allocated. We therefore have a vocabulary $\mathcal{V} = \{1, 2, \ldots, 50\}$, where the word 1 corresponds to the shortest time bin and the word 50 corresponds to the longest time bin. A document $d$ then consists of a sequence of $N_d$ words drawn from $\mathcal{V}$.

We train a separate LDA model on each of the two corpora independently. Each model is trained on all documents in the corpus in order to infer the underlying distributions. In particular, the inference identifies latent topics that describe documents, where each document can be thought of as a random mixture over topics; for document $d$, the probability distribution $\Theta_d(t)$ gives the probability of drawing a word from topic $t$. Each topic will itself consist of a distribution over words; for topic $t$, the probability distribution $\Phi_t(w)$ gives the probability of drawing word $w \in \mathcal{V}$ from the topic $t$. Note that since we are applying LDA to the two groups independently, we will discover different distributions for the control and Parkinson's corpora, which we label with the appropriate superscript. We instruct LDA to discover $T = 50$ topics for each corpus using $\alpha = 50/T = 1$ and $\beta = 0.1$ for the hyperparameters that describe the underlying prior Dirichlet distributions for $\Theta^{C,P}$ and $\Phi^{C,P}$ respectively. These values are chosen heuristically and follow the guidance given in Ref. [14]. The procedure of building LDA models from the datasets is illustrated in Figure 1.

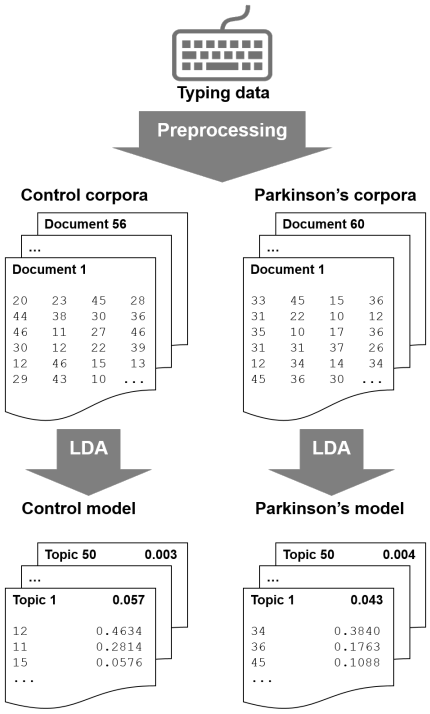


**Fig. 1.** An outline of the procedure: typing data is preprocessed to form a control corpus and a Parkinson's corpus. Each corpus consists of a set of documents (typing sessions) formed from a sequence of words (key stroke durations). LDA then infers 50 latent topics for each corpus.

# 4 Results

## 4.1 Topic discovery on Parkinson's and control groups

The LDA models trained on the control and Parkinson's groups are visualized in Figures 2 and 3. We emphasise that the words are common between the two datasets (i.e., $w = 23$, for example, will always refer to key presses of the same duration), whilst the meaning of the topics is different for the control and Parkinson's datasets (i.e., the control topic $t^C = 23$, for example, does *not* contain the same words as the Parkinson's topic $t^P = 23$). Figure 2 shows the composition of documents as a mixture of topics, and Figure 3 shows the word content of each of the topics.

Note that, within each corpus, topics are ordered according to their probabilities. Thus $t^{C,P} = 1$ corresponds to the topic that is most likely to be drawn for generating a document, $t^{C,P} = 2$ is the next most likely, and so on (the least likely topic is $t^{C,P} = 50$). The ordering of documents has no particular significance.

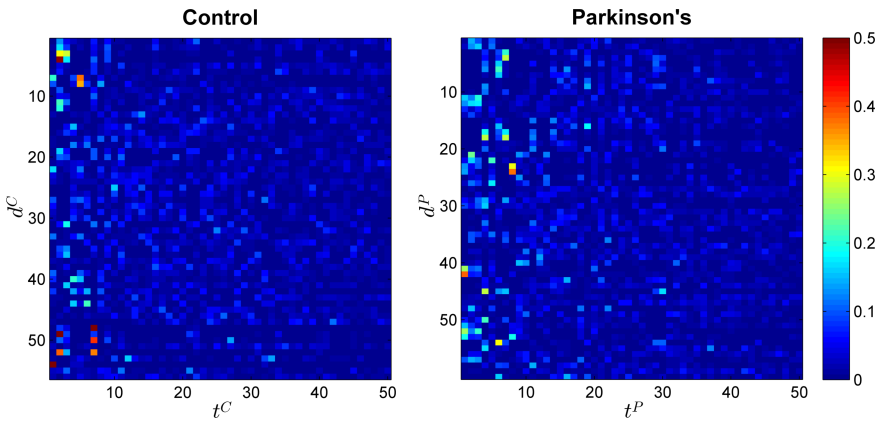


**Fig. 2.** The distribution of topics discovered in each document, $\Theta^C_{d^C}(t^C)$ and $\Theta^P_{d^P}(t^P)$. Results for the control corpus are shown on the left, and results for the Parkinson's corpus are shown on the right. The same colour scale has been used for both plots, and each row sums to 1. The points at $(t^C, d^C) = (7, 48)$ and $(1, 54)$ have been clipped to the limit of the colour scale.

Figure 2 shows that each document is indeed generated by a probabilistic mixture of a range of topics. As expected from the labelling, the most probable topics are towards the left of each plot. Although there are some particularly high density data points for the control corpus (indicating that a document is composed of just a few topics), there is no clear distinction between the control and Parkinson's corpora in terms of the distribution of topics over documents. In other words, both a control and Parkinson's typing session can be modelled using a similar distribution of topics. In order to discriminate between control and Parkinson's typing data, we turn to the content of these latent topics.

Figure 3 demonstrates that the words discovered in each topic are noticeably different for the control and Parkinson's corpora. There is a clear tendency towards Parkinson's topics containing words with higher labels, which correspond to longer key press durations. In particular, we might highlight $t^P = 4, 6$, which contain high density points around long key press durations, as particularly indicative. These topics are relatively likely to be drawn when generating a Parkinson's document, and no similar topics exist for the control corpus. They could thus be regarded as potential signatures for Parkinson's disease.

It is also worth noting that there appears to be some quite similar topics for the control and Parkinson's corpora, although the topic label (and hence relative probability) may be different, e.g., compare $t^C = 1$ with $t^P = 8$. In both corpora, a topic tends to be composed of words clustered around a certain label. This indicates that for both the control and Parkinson's subjects, similar length key press durations tend to co-occur in a given document. From our results the clearest signature of Parkinson's is the content of topics discovered by LDA;

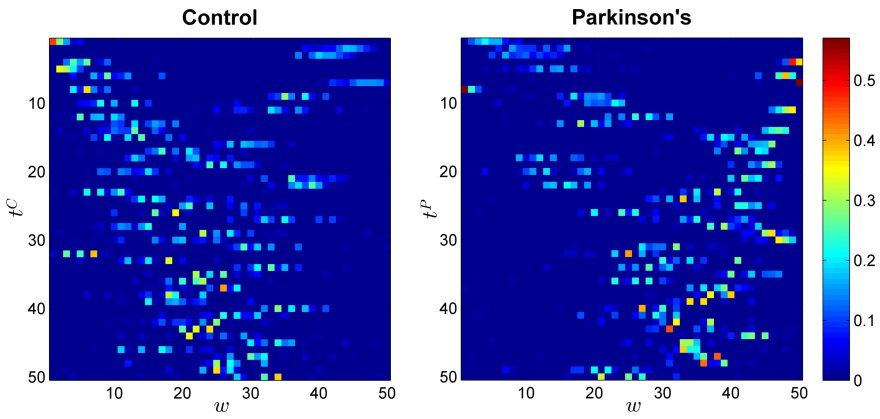


**Fig. 3.** The distribution of words discovered in each topic, $\Phi^C_{t^C}(w)$ and $\Phi^P_{t^P}(w)$. Results for the control corpus are shown on the left, and results for the Parkinson's corpus are shown on the right. The same colour scale has been used for both plots, and each row sums to 1. The point at $(w, t^P) = (50, 7)$ has been clipped to the limit of the colour scale.

in particular, topics which show long key press durations seem to be strongly indicative of Parkinson's disease.

### 4.2 Topic similarity analysis to discover Parkinson's behaviour

We now systematise the detection of signature Parkinson's topics by computing the similarity between topics discovered for the control group and those for the Parkinson's group. In particular, for each control topic $t^C$ and Parkinson's topic $t^P$ we compute the Bhattacharyya coefficient

$$\rho(t^C, t^P) = \sum_{w \in \mathcal{V}} \sqrt{\Phi^C_{t^C}(w)\, \Phi^P_{t^P}(w)}, \tag{1}$$

where the summation runs over all words in the vocabulary. This gives a measure of the overlap between the distribution $\Phi^C_{t^C}$ and $\Phi^C_{t^C}$. As a measure of the "uniqueness" of a given Parkinson's topic $t^P$, we then compute the average Bhattacharyya coefficient,

$$\bar{\rho}(t^P) = \frac{1}{T} \sum_{t^C=1}^{T} \rho(t^C, t^P). \tag{2}$$

A small value of $\bar{\rho}(t^P)$ indicates that there is on average little overlap between the content of topic $t^P$ and the control topics, and hence that $t^P$ is a signature of Parkinson's disease.

Figure 4 shows the Bhattacharyya coefficient for each pairing $(t^C, t^P)$ and the average Bhattacharyya coefficient $\bar{\rho}(t^P)$. Identifying the Parkinson's topics with

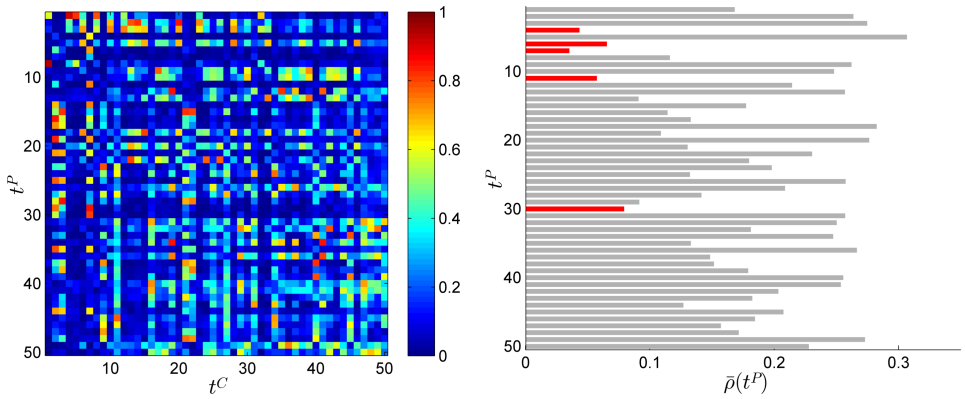


**Fig. 4.** Identification of signature Parkinson's topics. Left: the Bhattacharyya coefficient $\rho(t^C, t^P)$ gives a measure of the similarity between a control topic and a Parkinson's topic. Right: the average Bhattacharyya coefficient $\bar{\rho}(t^P)$ indicates which Parkinson's topics are on average the most dissimilar from the control topics. We have highlighted the five topics with the lowest $\bar{\rho}(t^P)$; these topics are signatures for Parkinson's disease.

the smallest $\bar{\rho}(t^P)$ confirms our above interpretation of $t^P = 4, 6$ as signature topics for Parkinson's disease and also identifies $t^P = 7, 11, 30$. The content of the topics is shown in Figure 5, which indicates that they are dominated by words that correspond to long key press durations.

| $\boldsymbol{t^P = 4}$ | | $\boldsymbol{t^P = 6}$ | | $\boldsymbol{t^P = 7}$ | | $\boldsymbol{t^P = 11}$ | | $\boldsymbol{t^P = 30}$ | |
|---|---|---|---|---|---|---|---|---|---|
| $w$ | $\Phi^P_{tP}(w)$ | $w$ | $\Phi^P_{tP}(w)$ | $w$ | $\Phi^P_{tP}(w)$ | $w$ | $\Phi^P_{tP}(w)$ | $w$ | $\Phi^P_{tP}(w)$ |
| 49 | 0.483 | 48 | 0.390 | 50 | 0.964 | 49 | 0.365 | 47 | 0.358 |
| 50 | 0.369 | 47 | 0.243 | 44 | 0.013 | 48 | 0.294 | 48 | 0.277 |
| 48 | 0.108 | 49 | 0.230 | 46 | 0.008 | 47 | 0.201 | 49 | 0.187 |
| 47 | 0.033 | 44 | 0.115 | 42 | 0.007 | 46 | 0.126 | 44 | 0.065 |
| 44 | 0.003 | 40 | 0.017 | 28 | 0.003 | 39 | 0.008 | 45 | 0.062 |

**Fig. 5.** The content of the topics discovered that signify Parkinson's disease. For each topic the most likely 5 words are shown, together with the corresponding probabilities.

### 4.3 Discussion

In preparing this paper we also considered a number of features which are not discussed here, including the travel time between key strokes as well as the hand

(left vs. right). We found the key press duration time to reveal the most interesting differences between the two groups. However, considering more features related to typing activity such as key pressed, keyboard row, or even smartphone key and holding information, is an avenue for further work.

These initial results are promising and the approach can be generalised to other datasets and applications. One possible limitation with the approach is that the topics found across the two corpora are different, although the words are consistent. Our initial experiments learned a topic model on both corpora combined, but the topics discovered showed a mixture of behaviours and the results did not address the task well. In future work, extensions to the graphical model of LDA to learn consistent topics across two groups without combining the data or learning them completely separately will be considered.

## 5 Conclusions

In this paper, we aimed to find the differences in typing behaviour between people who have Parkinson's disease and those who do not. Considering a dataset of 116 typing sessions each of 15 minutes duration, we formulated an approach based on topic models to identify the patterns that are much more probable in the group with Parkinson's disease than in the control group. These patterns corresponded to longer key hold times. The novelty of this work stems from the ability to display the word distributions from topics which correspond to the actual behaviour of interest. This is particularly useful when considering more complex typing features, as well as multi-modal typing features (e.g., hand, key and hold time).

Several future directions arise based on this work. The most elementary of these is to consider more features and fuse them appropriately in order to obtain more intricate behaviour differences between the two groups. Experiments on other datasets to validate the generalisation of the results would also be important. We also plan to utilise generative models that learn keystroke dynamics and further evaluate the discovered topics in order to detect signatures of Parkinson's disease in key stroke logs.

## References


1. Bachlin, M., Plotnik, M., Roggen, D., Maidan, I., Hausdorff, J.M., Giladi, N., Troster, G.: Wearable assistant for parkinsons disease patients with the freezing of gait symptom. IEEE Transactions on Information Technology in Biomedicine 14(2), 436–446 (2010)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research 3(Jan), 993–1022 (2003)
3. Chen, B.R., Patel, S., Buckley, T., Rednic, R., McClure, D.J., Shih, L., Tarsy, D., Welsh, M., Bonato, P.: A web-based system for home monitoring of patients with parkinson's disease using wearable sensors. IEEE Transactions on Biomedical Engineering 58(3), 831–836 (2011)

4. Farrahi, K., Gatica-Perez, D.: Discovering routines from large-scale human locations using probabilistic topic models. ACM Trans. Intell. Syst. Technol. 2(1), 3:1–3:27 (Jan 2011), http://doi.acm.org/10.1145/1889681.1889684
5. Giancardo, L., Sanchez-Ferro, A., Arroyo-Gallego, T., Butterworth, I., Mendoza, C.S., Montero, P., Matarazzo, M., Obeso, J.A., Gray, M.L., Estépar, R.S.J.: Computer keyboard interaction as an indicator of early parkinsons disease. Scientific reports 6, 34468 (2016)
6. Hazan, H., Hilu, D., Manevitz, L., Ramig, L.O., Sapir, S.: Early diagnosis of parkinson's disease via machine learning on speech data. In: Electrical & Electronics Engineers in Israel (IEEEI), 2012 IEEE 27th Convention of. pp. 1–4. IEEE (2012)
7. Huynh, T., Fritz, M., Schiele, B.: Discovery of activity patterns using topic models. In: Proceedings of the 10th international conference on Ubiquitous computing. pp. 10–19. ACM (2008)
8. Lin, C., He, Y.: Joint sentiment/topic model for sentiment analysis. In: Proceedings of the 18th ACM conference on Information and knowledge management. pp. 375–384. ACM (2009)
9. Madan, A., Farrahi, K., Gatica-Perez, D., Pentland, A.S.: Pervasive sensing to model political opinions in face-to-face networks. In: International Conference on Pervasive Computing. pp. 214–231. Springer (2011)
10. Mariani, B., Jiménez, M.C., Vingerhoets, F.J., Aminian, K.: On-shoe wearable sensors for gait and turning assessment of patients with parkinson's disease. IEEE transactions on biomedical engineering 60(1), 155–158 (2013)
11. Mei, Q., Ling, X., Wondra, M., Su, H., Zhai, C.: Topic sentiment mixture: modeling facets and opinions in weblogs. In: Proceedings of the 16th international conference on World Wide Web. pp. 171–180. ACM (2007)
12. Patel, S., Lorincz, K., Hughes, R., Huggins, N., Growdon, J., Standaert, D., Akay, M., Dy, J., Welsh, M., Bonato, P.: Monitoring motor fluctuations in patients with parkinson's disease using wearable sensors. IEEE transactions on information technology in biomedicine 13(6), 864–873 (2009)
13. Salarian, A., Russmann, H., Wider, C., Burkhard, P.R., Vingerhoets, F.J., Aminian, K.: Quantification of tremor and bradykinesia in parkinson's disease using a novel ambulatory monitoring system. IEEE Transactions on Biomedical Engineering 54(2), 313–322 (2007)
14. Steyvers, M., Griffiths, T.: Probabilistic topic models. Handbook of latent semantic analysis 427(7), 424–440 (2007)
15. Tsanas, A.: Accurate telemonitoring of parkinsons disease symptom severity using nonlinear speech signal processing and statistical machine learning. Diss. University of Oxford (2012)