Similarity

All UCAS Undergraduate and UCAS Conservatoire personal statements are screened by our similarity detection system, Copycatch. Each personal statement is compared against a library already in the UCAS system, and a library of sample statements collected from a variety of websites and other sources, including paper publications. After it has been processed, each new personal statement is added to the library.

Verification Team process

Any statements showing a potential level of similarity of greater than 30% will be reviewed by members of the Verification Team and either cleared or marked as containing similarity. The Verification Team also ensures personal statements have not matched with a previous personal statement from the same applicant.

For those personal statements considered to contain a sufficient degree of similarity to the matched source(s), automatic emails are generated to:

- the applicant – referring them to My Application, where they will be able to view a copy of the colour-coded transcript of their personal statement
- the applicant's choice(s) – with a link to the colour-coded transcript

A similarity detection service report will be emailed to the Similarity Contact at all course providers listed in the application. This email will contain a link to the colour-coded transcript of the personal statement, where there are reasonable grounds to suspect collusion. The report is designed to provide admissions staff and decision-makers with additional information to be considered when making decisions about applicants.

The decision about what action, if any, to take regarding notified cases rests with individual providers. UCAS is not involved in the decision-making process.

The principles behind Copycatch

In English texts of any sort, there is a division between the grammar words and the content words. Grammar words are frequently used throughout a text to organise the content for a reader. The most common of these grammar words are **the** and **of**, which by themselves account for approximately 10% of all words used in English texts. The Copycatch process has a function that excludes approximately 450 of these words from the detection process, as all personal statements would reasonably be expected to contain them.

Content words make up approximately 50% of any text. The majority of these words will appear only once, with a small number appearing twice. Copycatch analyses these content words to identify combinations and sequences common to more than

one personal statement. Normally picking any two independently prepared personal statements at random, the amount of overlap between the infrequently used content words will be very small. This is what we mean by using your own words, and is the result of having different experiences, different wider vocabulary, and different ways of expressing yourself.

There are a number of content words that will reasonably occur in a large number of statements. For example: Duke, Edinburgh, gold, silver, bronze, football, netball, rugby, swimming – to name a few. These words are also excluded from the similarity detection process.

What is checked

We have a library of approximately 1,500,000 personal statements to compare new applications against. This library is continually added to with new personal statements, as they are received. We also have a library of approximately 600 example personal statements, from a variety of websites.

What checking is done?

Copycatch analyses personal statements sentence by sentence. Complete copying of statements is rare, and copying part or all of a sentence is more common. Copied sentences are often modified, as the example shown later demonstrates. Copycatch is able to identify sentences where this might have taken place, and indicates this on the report for consideration by the provider.

What do we count as a sentence?

A core function of Copycatch is the identification of where a sentence starts and finishes. The main principle is to use a full stop, exclamation mark, or question mark followed by a space as a sentence delimiter. The refined process is more sophisticated, also recognising quote marks, brackets, etc. It also recognises that the above rules don't always apply, as in **etc.**, for example, and where applicants leave out the space after the full stop because they are restricted to exactly 4,000 characters in their personal statement. Others omit full stops and just use a capital letter to indicate a new sentence. This can be intentional or unintentional, so Copycatch checks for the following patterns '...riding an **elephant.That** is why...'; and '...driving a tractor **So** I became...'. It treats both the **t.T** and **r S** as sentence boundaries.

The numbers involved

Copycatch compares each sentence of every personal statement with all the others we hold and, therefore, the numbers are huge. The average personal statement contains 30 sentences, so there are 15 million of them to check against each year, in addition to those collected from web sources and books.

Filter settings

Copycatch uses a set of filters to ensure the sentences identified are the closest match to those in the applicant's personal statement. The process then checks there is sufficient inherent similarity in these sentences for them to be identified as potentially copied. Finally, it checks there are a significant number of potentially copied sentences in the statement.

The matching process

Although some applicants copy whole sentences, paragraphs, or complete personal statements, the most common form of copying involves some modification of the sentences used. This may be simply changing the course name or subjects studied, or may involve more extensive insertion or deletion of material. Copycatch is designed to look for partial matching in these situations, enabling it to identify changes in word order. All comparison between sentences takes account of the context in which the words appear. Only words that appear reasonably close together in both sentences are paired. This avoids arbitrary matching of sentences which happen to contain only a few of the same words.

Copycatch produces a marked-up copy of the statement as shown below. Where a significant number of sentences have been identified from one source the marked-up report will state for example:

'These 7 sentences were found in a personal statement submitted on Apr 24, 2018'

Or, if the source has been identified on the internet:

'These 6 sentences were found at www.thestudentroom.co.uk on Aug 16, 2016'

If Copycatch identified possible copying from a number of sources, it will report:

'4 further sentences were found in 3 files'.

Example personal statement

I have always been fascinated by the way writers can influence and even manipulate reader's emotions by their expression of thoughts and by their ability to encourage the expansion of our imaginations and understanding. My favourite authors include Phillip Pullman and Caroline B. Cooney whose novels are inspiring because of their enviable lucidity and innovative character development.

I had a vivid imagination as a child possibly influenced by my interest in the captivating work of such authors as Enid Blyton, Roald Dahl and Charles Kingsley. I still enjoy reading in my spare time. It is impossible for me to choose my favourite book but I have taken pleasure numerous times in reading 'A Child Called It' by Dave Pelzer, a book which I find mesmerising and deeply moving and 'Little Women' by Louisa May Alcott, a powerful and inspiring novel. I enjoy literature that can provoke a range of emotions in the reader from start to finish and I think that these books fit the bill perfectly. What attracts me most to English is not only the chance to expand my literary knowledge, but the opportunity I am given to communicate my own thoughts and ideas. I love how writers can express their opinions and emotions in words that can influence, inspire and touch others. For this reason, creative writing is one of my favourite aspects of the subject. The way I can express myself freely on paper is a liberating feeling for me and I find writing to be a useful and therapeutic way of conveying my emotions.

What the sentence colours mean

Copycatch marks up text with red, blue, and pink where exact matches between the personal statement and information held in the library have been identified.

Words shown in black are words in a potentially copied sentence that are not contained in the library version of that sentence.

Words in black and underlined are words in a potentially copied sentence that are not identical to the library version of the sentence, but are contextually similar. For example:

'a powerful and inspiring novel' in the text above might read 'a powerful and inspiring book' in the library text.

Red is used for sentences from the most matched personal statement. You can see that there are five in this extract alone, four of which are exact matches. The first one shows **Enid Blyton** and **Charles Kingsley** in black, meaning that different authors are in this position in the matched sentence.

Blue is used for the next best matched statement. Only one is shown here and this is a complete match of a sentence from a web source.

Pink is used for the third best match. Two of them are shown in this extract. The first has two underlined words **reader's** and **understanding**, indicating that a similar form of the word has been used in the matching sentence. This could mean that this writer has added the apostrophe in **reader's** and rephrased the way **understanding** was expressed in the matched sentence.

Brown is used for all other matched text, regardless of source.

Grey is used for sentences for which no match has been found, and for very short sentences which don't get checked.

The similarity calculation

The similarity percentage is calculated by dividing the number of matched sentences by the total number of sentences. This means that sentences which match exactly and those which match partially are treated the same way. So, a 100% match may show sentences which are completely identical, or have some alteration in them which meets the sentence similarity criterion.

The dates on the matched personal statements

At the end of the marked-up personal statement, the number of sentences matched to library or internet-based sources is shown in the same colour as that used to mark-up the sentences. The date is merely indicative of how long this personal statement has been in the UCAS collection. It does not mean this particular statement was the one used as the source for the current personal statement. Both may be taken from a source outside the library, or there may be other related files inside the library which have not been shown because there was no additional matched information.

The dates on the matching web sources

The number of web source sentences is shown in the same way, but here the date means either the date it was posted to the website, if known, or the date when the web source was identified by us. Again, it does not necessarily mean the file was the actual source. Some web sources are very popular and may appear on more than one website, or have been used in a modified form in a personal statement within the UCAS collection.

Can an applicant replace their personal statement?

UCAS will not accept any amendments to an applicant's personal statement after the application has been submitted. We advise applicants to contact their course providers directly if they have any additional information which they wish to be considered.

Can the applicant appeal?

Against the colour-coded transcript:

The transcript only highlights similarities to other data already available to us, and we do not envisage, at this point in time, any grounds for an appeal against this.

To the provider:

As with any application, you may receive a request for feedback on why you have made a particular decision. Any appeals should be dealt with in accordance with your own admissions policies and procedures.

Contact details

If you have any queries about the similarity detection service, please contact the Verification Team on similaritydetection.hei@ucas.ac.uk, or at 01242 545 494.