	Drop	Change	Other
Novel-Copy	24	102	0
Novel-Non-Copy	14	128	30
Observed-Copy	31	11	126
Observed-Non-Copy	12	1	186

Table 3: Counts of automatically detected output categories (*drop*, *change*, and *other*) for a sample of NNP tokens (EN–DE) that were not copied.

more or less copy-prone than expected. In both cases, the cause appears the same: the context occurred repeatedly in many very similar sentences in the training data. Highly copy-prone contexts that produced copying percentages greater than 70% even in observed-non-copy tokens often appeared in common boilerplate text (e.g. "stay at [NNP]" or "rates for [NNP]" followed by "Hotel").[15] Where we observe lower than expected rates (e.g. ") of [NNP]"), we find that the system may have memorized training sentences.

5.2 Analysis of Words That Are Not Copied

When words are not copied, what sort of output is the system producing? We find that it typically falls into one of four categories: *drop* (no target token aligns with the source token), *change* (the word is changed: partially translated, transliterated, or inflected even if it is not a target language word), *substitution* (the word is replaced with a fluent but not adequate substitute), or *translation* (translated into a target language word).

We begin with an automatic analysis. We randomly sample 200 examples each of sentences containing words that were not copied for novel-copy, novel-non-copy, observed-copy, and observed-non-copy NNPs (EN–DE). We retranslate each sentence and produce a soft alignment matrix from the attention mechanism, then convert the soft alignments between BPE segments into hard alignments between the source word and one or more target words.[16] A word has been *dropped* if it is unaligned. We count a word as being *changed* if any words it is aligned to have any subword (BPE segment) overlap with the original word's subwords. Both *substitution* and *translation* fall under *other*; we analyze those manually.

Results are shown in Table 3.[17] For all novel words, the most frequent output type is *change*. For example, the novel NNP *Bishnu* is changed into *Bischnu* in German.[18] Other changes include translations of parts of the word, and concatenation with other tokens. The output token often starts with the same character or sequence of characters as the source token.[19]

We manually inspect examples in the *other* category. For observed-non-copy words, almost all are translations (e.g. *Sea* translated correctly as *Meer*), as expected. For observed-copy words, we see a mix of translations and other changes to the words, which are almost evenly split between substitutions and small changes. These include inflections (e.g. *Bremen magazine* reasonably translated as *Bremer Magazin*[20]).

Within the *other* category, perhaps the most interesting cases are those where words appear to be substituted with a fluent but not adequate alternative. Many substitutions occur when the rare word is inserted next to a word that often forms a collocation (like "United States" – in sentences that include "in the [NNP] States" the translation sometimes defaults to a translation of "United States" regardless of the actual NNP inserted in place of "United"). Others have a less common NNP swapped for one that belongs to a similar semantic category (e.g. the place name *Dublin* being generated instead of the less common *Halle* – as Arthur et al. (2016) and others observed). For novel-copy words labeled as *other*, three quarters are substitutions and one quarter exhibit small changes. The reverse is true for novel-non-copy words: the majority exhibit small changes while almost thirty percent are substitutions.

5.3 Properties of Copied Words

Certain words exhibit properties that make them more likely to be copied, regardless of context. At first glance, it seems unintuitive that the rate of copying of novel-copy words and novel-non-copy words differs (Fig. 1) – the model has never observed any of these words, and they are being presented in identical contexts – why does it differentiate between them? Doing so indicates that the model has learned what makes a sequence of

[15] Since hidden representations contain whole sentence information, right side context may influence copying too.

[16] We use AmuNMT (Junczys-Dowmunt et al., 2016), producing slightly different output. See Appendix C for details.

[17] Rows do not sum to 200 because some words in our random sample were copied by the the AmuNMT decoder.

[18] A near-transliteration – the "sh"/"sch" transformation is seen in EN–DE cognates, e.g. "ship" and "Schiff".

[19] Appendix D contains examples of this and more.

[20] *Bremen* and *Bremer* are unique BPE segments, so the *change* heuristic could not be applied.