1. Определить алфавит, т.е. символы, которые могут образовывать лексемы. В случае русского языка - это русские буквы, цифры, латинские буквы и специальные.
2. Сокращение – тыс., кв.
3. Аббревиатуры
4. Числа
5. Даты
6. Инициалы
7. почтовые адреса
8. url адреса

Варианты написания чисел – 100

100.1

Для начала определяем предложения так как будто предложения были написаны корректно. Т.е. справа на лево ищем знаки пунктуации: “.”, “!”, “?”, “…”. Все то чем может оканчиваться предложение. Определенное прежложение записывается в массив и в массиве отмечаются уже найденные характеристики этого предложения. Например, предложение оканчивающееся на “.” – получает характеристику –

В начале возможно преобразование текста, для уменьшения количества символов – удаление лишних «мягких» символов – задвоенных пробелов, табуляций, символов возврата каретки…

Для простоты – все предложения оканчиваются знаком пунктуации.

1. Определяем цифровые последовательности;
2. Определяем цифровые выражения (например, 2+x=3 или x>5);
3. Удаление/замена «ненужных», неинформативных символов:
   1. последовательности непечатаемых символов (пробелов, табуляций, переводов формата, переводов строки, возвратов каретки);
   2. ведущие и замыкающие непечатаемые символы;
   3. непечатаемые символы после символов словарей, но перед знаками пунктуации – «,», «.», «!», «?», «;», «»», «%»(для цифровых последовательностей), «:», «\*» (для цифровых последовательностей), «’», «)», «}», «]»;
   4. непечатаемые символы после знаков пунктуации, но до символов словарей – «-« (не для цифровых последовательностей), «»», «’», «(», «{», «[»;
4. Замена задублированных знаков пунктуации одним единственным;
5. Замена последовательности символов «.», которых больше 2 на «…» - одиночный символ;
6. Замена последовательности символов «!!!?!!!» или «?!» на «¡» (u00a1);
7. Замена последовательности (?) и (!) на « «;
8. Вставка после символов пунктуации символа пробела;