Aligning and patterning patent quality

Gabriel Luis-Franchy (University Graz),

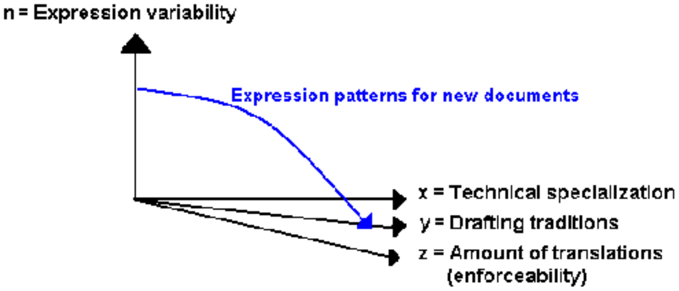
Within the frame of the European Patent Convention (EPC in the following) various approaches have been described in order to reduce patenting costs, especially for innovative SMEs, and to improve patent quality for the benefit of society as a whole. The main focus of these approaches has been laid on lowering procedural costs by reducing translation requirements; other potential cost-saving measures have been neglected, mainly due to political reasons. Following paper gives a first insight into the possibilities of an integrated documentation approach in order to streamline the patenting procedure using lexicographic research in a semantic aligning environment.

Legal background

The legal workframe in Europe features a central entity, the European Patent Office (EPO in the following), that grants european patents as executing body of the EPC. In a relevant number of EPC contracting countries, the preliminary european patent must be fully translated in order to be published in each national industrial property bulleting, (validation) being afterwards enforceable in that particular country. Much effort has been put in order to streamline the validation process, specially through the London Protocol, but little improvement from the economic point of view has been reached. Obviously, for a relevant number of EPC-countries it is a key concern to have IP-certificates fully, e.g., accurately translated in their national language, since the interpretation of national law, as well as the interpretation of national identity, is strongly tied to national language. Unless a fair validation option is approved that takes fairly into account the language issue, every new draft of European patent law is exposed to structural failure.

Linguistic Background

There is a high level of consent in linguistics about the relationship between specialized languages and expression variability, i.e., the more specialized the language filed, the less variability it allows when rendering its textual objects. Patents feature one of the highest specialization levels in technical literature (Schamlu 1984). It is also proof available about the fact that, according to this high specialization level in patent talk, the variability in patent language can be subsumed (Göpferich 1992) to an ideal document structure (macrostructure). The following graphic illustrates the relationship between specialization and expression variability:


n = Expression variability
Expression patterns for new documents
x = Technical specialization
y = Drafting traditions
z = Amount of translations
(enforceability)

Picture 1: Relationship between variability and high specialization

The global legal character of EP-PCT patents shows also that quality patents are often translated into various languages. In a similar way as literature, the amount of translations can be seen as one yardstick to measure the worldwide resistance of a quality patent. Obviously this yardstick must be combined with other parameters for measuring patent quality, such as compliance with legal frameworks (WIPO1

Standard 36). The amount of translations is not only a potential yardstick for measuring patent quality, but also one of the biggest problems the EPC is facing, since translation requirements and/or costs within the EPC-frame seem to be the main obstacle in order to ensure patent quality in Europe. The EMTP program started by the European Patent Office in collaboration with the American language provider Wordlingo represents a try to help companies and the public to gain insights into the content of patents using machine translation yielding limited results. The limitations of machine translation are obvious, specially when the machine has to deal with complex, unrestricted syntactic structures. Therefore, according to Mitamura (1998) and Gajer (2008) high-quality machine translation between human languages for unrestricted text is still a long-term scientific dream, which is especially hard to become true in the technical-legal discourse. Back to the realm of reality and since the beginning of the 90´s the raise of computer aided translation (CAT in the following) has shown great results.

Technical background

The introduction of multilingual data description formats compatible with the eXtensible Markup Language (XML) has leveraged the quality, consistency and the speed of translation processes. The standard in this field, the translation memory exchange format (TMX), keeps spreading and has even been adopted by the EU-Commission to make both public and reusable the records of the Acquis Communitaire, among other legal texts. In the present it is possible to create TMX-files using two different techniques:

- CAT- translation: The user (human translator) renders the translation and saves it in this format
- CAT- aligning: The user correlates two (or more) input texts, depending on their availability/quality

In the research underlying to this paper, CAT-aligning has been applied to the field of patent literature. This approach has shown a speed rate of 1:12, providing 100% output quality in fully parallel texts. Non parallel texts ("noisy input" in the following) poses a problem since the existing input material (PCT and/or EP patents) often feature elements that are not present in the original ("additive noise") and/or elements left out from the original ("negative noise"). Since the computer can not establish a 100% valid correlation in a noisy environment, it is necessary to apply a more flexible alignment method that allows both the creation of highest quality TMX output as well as the semantic description of noisy-phenomena. Although aligning is a simple and fast method to create information units in TMX-format, it provides some limitations when it comes to describe features below (syntactic) and above (pragmatic) the sentence level. The development carried out within the Patterm project provides an alignment environment which allows for aligning as much texts as available, featuring segment tagging both on the semantic and the pragmatic level. Following this approach, the pragmatic noise of patent literature presents no obstacles for aligning tools since the different claiming and/or descripting traditions of each EPC-PCT cultures render noisy input that can now be aligned semantically. According to this, our semantic aligning environment provide features to

- establish a multilingual semantic correlation on sentence level, similar to TMX;
- describe noise phenomena such as claim, description and system shifts and
- detect patterns of inventiveness scoring the segments-documents with highest compliance to patent xml/xsd schemata, highest amount of translations and lowest levels of data noise.

Obviously these features can only be implemented using human editing, although automation of key- workflow steps can be easily put into practice. The potential losses of output speed can be neutralized by the virtualization of the aligning environment. The advantages of this approach are numerous: On the one hand the semantic aligned output allows for information retrieval for translation processes, on the other hand, according to the linguistic axiom of high specialization vs. low variation it also allows for the authoring of new patent documents reusing the patterns of inventiveness harvested during the semantic aligning.