Language diversity and implications for Language technology in the Multilingual Europe

Cristina Vertan
University of Hamburg
cristina.vertan@uni-hamburg.de

Walther v. Hahn
University of Hamburg
vhahn@informatik.uni-hamburg.de

Europe has a particular and unique setting. On one had it has a great language diversity, there are twenty four official languages and a dozen of minority languages largely used. On the other hand most of these languages belong to one of the indo-European language families (Roman, Germanic Slavic) and within these language families similarities at lexical and syntactic level can be observed. Whilst an increased attention have been given to the development of language technology tools for the official EU –Languages, processing tools for minority languages have a chance to progress only by exploiting similarities within their language families.

In order to have an overview about the European linguistic diversity and the implications on the language technology research we republish here a part of the article

"Translation Difficulties and Information Processing Problems with Eastern European Languages"

Cristina Vertan and Walther v.Hahn
Published din the volume "Multilingual Processing in Eastern and Southern EU Languages", Cambridge Scholar Press, 2012

It is still popular today to blame machine translation (MT) for poor translations of literary texts. However, even inelegant translations are an industrial factor in producing MT software, in selling multilingual retrieval for relevance scanning or in opening markets by issuing simple foreign-language descriptions. Information retrieval (IR) technologies are effective even if their degree of linguistic correctness is low.
The success story of Machine Translation is partly owed to some simplifications, which made its start-up easier (leaving aside the political presetting of English-Russian translations). Simplifying reality was a promising approach, because the reduction of parameters from syntax, morphology, and domain coverage formed the basis for the demonstration of MT's feasibility.
Moreover, the statistical approach in MT nourished the hope that reasonable results for English can be seen as evidence for the fact that MT can be done with similar quality for any other language.
In subsequent decades experiments were performed with numerous other language pairs around the world including languages even, for which detailed linguistic knowledge was unavailable . The goal of these scientific and industrial research efforts was mainly to estimate the quality and costs of acceptable MT products for the commercially meaningful language pairs. The same holds true for multilingual information retrieval and multilingual information processing on other fields.
With the ever growing number of language pairs for which customers require cost-efficient processing, four aspects became clear:
1. There are domains and language pairs for which not even human translation/IR is available, e.g., financial law texts from Finish to, say, Hausa. The question remains how to obtain these at all.
2. There is no representative bilingual data collection (a "corpus") for these language pairs at all. Statistical approaches hence will not be feasible within the next 5-10 years. How to obtain inexpensive translations in the mid-term for these "low resourced" languages?
3. Many languages (e.g., Hausa) have more than one writing system or changing orthography (compare the post-reform rules for German orthogography that are in force since 2006). This poses the challenge of how to obtain homogeneous corpora.
4. Multinational or global companies need language processing for promotion, local instruction, or contracts, that affords legally binding results.

Proceedings of the Adaptation of Language Resources and Tools for Closely Related Languages and Language Variants, pages 2–6,
Hissar, Bulgaria, 13 September 2013.

The optimism of the pioneering years has yielded to scepticism regarding general recipes for multilingual processing such as translation, even for the traditional Western languages. In Europe, the expansion of the EU additionally demonstrated that democratic co-operation requires a huge work load of translation and bilingual information processing among today's 23 official languages. The sheer number of languages, their diverse linguistic structure and their different public use are reasons enough to give up some of the starting assumptions and simplifications of the first decades.
We discusses the rather different situations in Europe with regards to cross-lingual processing tasks in an English and American context along the following dimensions:

1 Languages

There are 230 spoken languages in Europe. Most of them have a long common history in the Indo-European paradigm. Even among the 23 official languages of the EU there are the Finno-Ugric official languages Finnish, Estonian and Hungarian. The Turkic and Mongolic families also have several European members, while the north Caucasian and Kartvelian families are important in the south-eastern border of geographical Europe. The Basque language of the Western Pyrenees is an isolated language, unrelated to any other language group in Europe. Much less known even to Central European citizens is the existence of a European Semitic language, the Maltese, written in Latin letters.
In the current volume we decided to refer only to the official EU languages, as representatives of most of the families enumerated above. Additionally, due to European integration there is an increased need in translation and cross-lingual management of documents in these languages. We hope that the some solutions presented here can be applicable also for non-official and minority languages of the EU.
Even the simple enumeration of the language families encountered in Europe already reveals the existence of major graphemic, phonetic and structural differences amongst them. The aim of this volume is not to investigate these differences from a linguistic point of view, but rather to insist on those discrepancies that trigger challenges for any translation system or cross-lingual/multilingual application. In this sense the following aspects are of relevance:

1.1 Writing differences

Although Europe has no unusual iconographic or syllabic writing systems but only phonographic paradigms, there are nevertheless problems with gathering homogenous bilingual language resources, i.e., training material for statistical approaches.

Cyrillic transliterations for named entities (NEs) follow four different (target language independent) transliteration schemata and numerous (target language dependent) transcriptions. As an example, consider the (operating system dependent) specific encodings for Bulgarian Cyrillics in contrast to Russian encoding. A similar situation exists for Arabic NEs in Maltese. The transliteration is not always standardised, which often leads to data sparseness. One word transliterated in three different ways will be identified in fact as three different words. Moreover, there is a problem with older electronic resources that were developed before the introduction of the Unicode character set: Many languages adopted transliteration simplifications that induce undesired ambiguities. For example the Romanian word for "goose" contains two diacritics "gâscă". A corpus collected before the introduction of the Unicode system would simplify this word to "gasca", which may be read also as "gaşcă" denoting a group of young people.

1.2 Variety of Linguistic Structure

European languages differ centrally in their use of pronouns and articles. So called "pro-drop" languages like Italian do not express the 1st person sg. pronouns explicitly, but mark them morphologically, whereas non-pro-drop languages like German have to add the pronoun explicitly in examples like "Ich gehe zu ihm" (I am going to him). Even more difficult in this sense is Hungarian, where lots of particles are only attached as morphemes (összerakhatatlanságukért = "for their quality of not being easy to put together")[1].
Grammatical gender is present or not (English, Basque), is expressed in noun endings (Italian, mostly), or not (German, mostly), affects other words by agreement (Spanish, French) affects the demonstratives (Italian), or not (Greek), is additionally marked by articles (German, but not

[1] This example is owed to Merényi Csaba from MorphoLogic

unambiguously), or not (Bulgarian, the Baltic languages). Articles are in use for the three-gender system (German), or two genders (Maltese), or the middle (Romanian, common singular for masc. and ntr.) attached to the end of a word (Romanian), or separated in front of the noun (French). Moreover, grammatical and natural gender have an unclear relation in most languages.

All these are major challenges especially in machine translation (MT) whenever the target language is more productive in pronouns or articles than the source. Rule based MT needs a deep linguistic analysis module and often the involvement of large knowledge bases in order to infer the correct target pronoun, while corpus-based MT cannot cope with this problem at all. In most cases the translation lacks not only the correct pronoun but also all derived information such as the correct inflection of the dependent nouns, adjectives and verbs.

To express definiteness, some languages use articles, while others express it by word order, which normally gets lost in surface-form statistical MT systems.

The word order of adjective and noun is semantically relevant in Spanish, restricted in German and fixed in English. Also the position of a verb in the sentence varies among language families. This is a real challenge not only for translation systems but also for multilingual tools that try to apply the same analysis technique to several structurally different languages. Rule-based tools lack a substantial number of common rules. Statistical methods, on the ther hand, require the availabilty of huge non-sparse data covering all these phenomena.

Word composition plays a major role in many EU languages and the order of components is significant. Sometimes logical particles must be inserted for correct translation. Distant verb particles in German are very difficult to differentiate from prepositions when only statistical methods are being applied. Again, such particularities constitute challenges not only in translation but for any preprocessing step in cross-linguistic processing.

1.3 Contact

All these languages have been in extensive contact with each other over time with the result, that additional irregularities were introduced. In Romanian, for example, one third of the vocabulary stems from Russian, Hungarian and German and has only been assimilated superficially. This means that the graphemic rules for Romanian are not homogenous. The contact, however, differs from language to language. Compare Romanian-Italian and Slovene-Italian contacts, e.g. whichdiffer with respect to the historical time, when the contact was established. Italy influenced Slovenia much more through Venetian than through modern Italian.

Borrowings were often done only partially so not all semantics is preserved. A special situation is encountered on the Balkan Peninsula where vocabulary related to food, e.g., old weapons or customs are usually shared by neighbouring communities but not necessarily by whole countries. For example, a lot of regional words in the North-Western part of Romania (Transylvania) are in common with Hungarian and German, but unknown in the Southern part of the country, which otherwise uses more words shared with Turkish and Bulgarian.

This constitutes a real challenge for modern retrieval systems which make use of ontologies. Building a language-independent ontology is an extremely difficult task, and even word-based semantic networks are highly problematic. A series of papers published over the last years report the difficulties in adapting the English Word-Net to each of the Balkan languages, and the challenge of homogenisation amongst these Word-Nets.

These considerations, however, touch upon cultural differences, that are addressed in the following paragraphs of this paper. We demonstrate with a few observations that behind the language differences in the EU there are many more cultural differences than between two regions of one country.

2 Text Structures, Forms and Formats

Two textual peculiarities of European publications are very confusing in corpora:

European texts often quote passages in a foreign language such as English or other European languages, because of a close contact with that language. Translated quotations from web pages, hence, are not always in the correct language, as an MT tool has been used.

A typical text form for an application, for an objection, for an expertise etc. differs not only lexically but also in style and form between European countries.

3 Cultures of Textuality

In Europe the influence of English varies from country to country. A comparison of German and French shows that an official inhibitive language policy influenced the borrowings to a high degree in France. Looking to Hungarian one can observe that the French policy more or less succeeded in most technical fields, whereas in Hungary two different medical nomenclatures exist that in practice and international information exchange conflict severely.

Irrespectively of a country's official language policy, the language policy of companies is also changing dramatically. A recent study observed that in Germany slightly more than 50% of all companies use German as the only business language, another 20% use German and English or only English, respectively. Less than 4% use other languages.

In general, technical fields in countries have a different degree of textualisation dependent on technologies, which have a higher or lower distance to text production and use, e.g. carpet weaving, vs. violin making vs. ecological food supplier vs. photo-copying or services.

Correspondingly, the reference to computerised texts, e.g. interactive web forms or download resources of public service differs very much between European countries, say, between Finland and Poland. While web forms must be explained even for language minorities, paper forms are issued by offices, where misunderstandings may be resolved in direct contact.

4 Commercial Market Value

The possible market value of multilingual or cross-lingual technologies clearly depends on the mere number of publications accessible and resulting from trade and industry. If you compare a Slavonic language minority like the Sorbs in Germany to Polish speakers in Poland, it becomes clear that the industrial expansion in Poland and exports abroad result in a disproportionately more extensive language and information contact than that for Sorbian speakers. Translating a handbook of nano technology into Polish makes more sense than translating it into Sorbian or publishing Shakespeare's works in Frisian, another German minority language.

5 Background and Perspectives

To come back to the main issue: Language technology in Europe is not an extension of known technologies to new languages, but a multidimensional challenge for science, technology and politics of quite another order of magnitude. It will bind research groups, translators, software companies and politicians for the next 50 years at least.

There is a widespread conception that

- the rapid development of the Internet,
- with new web services,
- the globalisation of the markets and
- the increase of online transactions

are the main factors driving international research in language technology.

This argument is, at least in a European context, only partially valid. In the era when Internet was in its infancy, and most part of the online information was exclusively distributed in English, the Directorate General of EU "Linguistic Applications" was already concerned with the additional languages from the countries willing to join the European Union.

"With the expected enlargement of the EU following the accession of up to ten Central and Eastern European countries (referred to as CEECs, which is the usual EU abbreviation), the translation complexity takes a quantum leap. The current EU languages (n.R. situation in 1998) (Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish and Swedish) can be translated in 110 language combinations, as each of the 11 languages can be translated into 10 other languages. With the addition of 10 new languages (Estonian, Latvian, Lithuanian, Polish, Czech, Slovak, Hungarian, Slovenian, Romanian and Bulgarian) the complexity goes up to 21 x 20 = 420 language combinations, but there is no obvious political or linguistic justification for changing the European Union's official policy of supporting multilingualism, which finds its expression in the MLIS programme, among others."

(Poul Andersen, DG XIII EU Representative, in an article "Translation Tools for the CEEC Candidates for EU Membership - an Overview", *Terminologie et Traduction* 1.1998, pp.140-166).

One can only speculate about the development in Europe in multilingual language technology without the political changes of 1989. The rule-based machine translation system Systran was functional, with reasonable performance for the EU-languages and for the requirements of that

time, namely translation of easy official documents.

We assume that the dramatic development of multilingual language technology in Europe was in fact driven by two forces: The new political context and the social impact of the Internet, rather than the economy.

We are also convinced that the European approach to multilingual language technology gave an impulse all around the globe to develop applications for various language communities: Recently systems for several Ethiopian languages appeared, a machine translation system for Quechua was presented (to quote only a few examples).

.