Procesamiento del lenguaje Natural, nº 41 (2008), pp. 113-120 recibido 7-05-2008; aceptado 16-06-2008

# Comparing languages from vocabulary growth to inflection paradigms: A study run on parallel corpora and multilingual lexicons

## *Comparando lenguas desde el léxico a paradigmas de flexión: un estudio sobre corpus paralelo y léxicos multilingües*

**Helena Blancafort**[1,2] **Claude de Loupy**[1,3]

[1]Syllabs
2 rue de Fontarabie
75020 Paris, France
{blancafort,loupy}@syllabs.com

[2]Universitat Pompeu Fabra
La Rambla, 30-32
08002 Barcelona, España

[3]Laboratoire Modyco
Université de Paris 10
200 av. de la République
92001 Nanterre, France

**Abstract:** In this paper we report on a corpora and lexical comparative study on how to compare the difficulties of five languages (English, German, Spanish, French and Italian) for morphosyntactic analysis and the development of lexicographic resources. Experiments were conducted on two different sets of multilingual parallel corpora and two different morphosyntactic lexicons per language. We measure and compare statistics on dynamic and static coverage, form-lemma and morphosyntactic ambiguities in the lexicon and the corpus. In addition to this, we use the lexicons to automatically generate inflection paradigms and calculate how many inflection paradigms are needed per language. Results show the difficulty of working with multilingual resources and parallel corpora and offer some surprising quantitative results on differences in languages.
**Keywords:** computational lexicography, morphosyntactic lexicons, computational morphology, inflection, multilingual parallel corpora, comparison of languages for NLP.

**Resumen:** En este artículo presentamos un estudio comparativo de corpus y de léxicos con el objetivo de comparar las dificultades que representan cinco lenguas (inglés, alemán, español, francés e italiano) para el análisis morfosintáctico y el desarrollo de recursos lexicográficos. Para ello hemos llevado a cabo varios experimentos utilizando dos corpus paralelos multilingües y dos léxicos morfosintácticos por lengua. Primero comparamos los resultados cuantitativos respecto a la cobertura dinámica y estática, y las ambigüedades morfosintácticas de los léxicos y corpus. Además, a partir de los léxicos hemos generado paradigmas de flexión para calcular cuántos son necesarios en cada lengua. Los resultados muestran la dificultad de trabajar con recursos multilingües y corpus paralelos. También ofrecen resultados cuantitativos sorprendentes respecto a las diferencias entre lenguas.
**Palabras clave:** lexicografía computacional, léxicos morfosintácticos, morfología computacional, flexión, corpus paralelos multilingües, comparación de lenguas para el PNL.

## *1 Introduction*

In recent years the number of multilingual data on the Web has been growing in leaps and bounds. As multilingual processing is gaining in importance, it is becoming urgent, for NLP purposes, to understand better the differences between languages. In this article we present current work on how to compare the difficulties of five languages (English, German, Spanish, French and Italian) for morphosyntactic analysis and the development of lexicographic resources.

It is known, e.g., that Latin languages have a richer verbal inflection than English and that German has a richer nominal inflection. Traditional morphological typological studies

ISSN: 1135-5948 © 2008 Sociedad Española para el procesamiento del Lenguaje Natural

already describe several linguistic phenomena for the comparison of languages, but don't provide any quantitative information about them.

In this paper we present a corpora and lexical comparative study conducted on two sets of multilingual parallel corpora, the JRC-Acquis (Steinberger et al., 2006) and the bible (Resnik et al., 1999) using two different morphosyntactic lexicons per language: MulText (Ide and Véronis, 1994) for each language, FreeLing (Atserias et al., 2006) for English, Spanish and Italian, Lefff (Sagot et al., 2006) for French, and Morphy (Lezius, 2000) for German. We measure and compare statistics on dynamic and static coverage, form lemma and morphosyntactic ambiguities in the lexicon and the corpus. In addition to this, we calculate how many inflection paradigms are needed to handle inflection of open class categories in each lexicon.

The paper is organized as follows: first, we give a short overview of the state of the art; in section three we describe the resources we used. Next, we report on vocabulary growth and coverage comparison. In section five we tackle the issues of morphosyntactic complexity, ambiguity and also compare inflection paradigms. Finally, we draw conclusions and discuss further work.

## *2 State of the art*

### 2.1 Comparing languages

Traditional typology distinguishes four types of languages: isolating, agglutinative, inflectional and polysynthetic. As observed by Trost (2003), this classification is quite artificial and real languages rarely fall into one of those classes: Chinese, e.g., is an isolating language but does have some suffixes. Pirkola (2001) expresses the need for a language typology for IR and suggests using the index of synthesis and fusion (Comrie, 1989; Whaley, 1997) to measure morphological phenomena. Furthermore, he suggests finer-grained indexes and semantic analysis. He claims that by combining these variables it would be possible to predict the performance of morphological processing and hence, the difficulties that a given language represents for IR.

### 2.2 Induction of morphological rules

Lexicographic resources are needed for basic morphosyntactic analysis like lemmatization. The difficulty and time needed for accomplishing these tasks depend on the characteristics of a language. Latin languages, e.g., are supposed to be longer to encode than English because of their verbal inflection paradigms. Hence, it is quite common to develop an inflection engine using hand encoded inflection rules. This can be a time consuming task for languages with rich inflection. In the case of Spanish, e.g., a verb paradigm can contain more than 40 forms. Besides, the number of inflections for a lemma is irregular, which implies that two verbs will not always have the same number of forms. Furthermore, there can be variants for the same inflectional form (e.g., two different participles such as *imprimido* and *impreso* in Spanish and also orthographic variants like the French verb forms *essaie* or *essaye*).

More recently some work has been carried out on automatic induction of morphology. Schone & Jurafsky (2001) designed an algorithm for inducing inflection rules in German, English and Dutch from a corpus without any human intervention. As far as we are aware, they have obtained the best results for a knowledge-free algorithm. Clément et al., (2004) present work carried out to build a French lexicon from a big corpus using morphological information. They apply a verbal inflection engine developed manually following the inflection patterns for open classes described in French grammars. We are not aware of any studies concerning the induction of inflection rules directly from a morphosyntactic lexicon. This is what we have carried out for the quantitative comparison of inflection paradigms (section 5.2).

## *3 Description of the resources: lexica and corpora*

### 3.1 Description of the lexica

To minimize the bias introduced by the lexicons, we used two different lexicons per language, the Multext and the FreeLing lexicons (v2.0). As FreeLing is not available for French and German, we took other large-coverage lexicons: the Lefff for French and Morphy for German.

One of the main goals of MulText was to develop monolingual and multilingual linguistic resources and to ensure the comparability and harmonization of tagsets in several European languages. Linguistic information is coded in a simple form-lemma-tag format. The tags are common to all languages. The definition of a tagset for all languages is not an obvious task, as there is an intrinsic incomparability of the tagsets due to the specifications of each language. Indeed, some tags are language specific. When this is the case, the attribute is marked with "-", as for Latin Languages having no case attribute.

Despite the big effort made for the harmonization of multilingual tagsets and lexical resources, the MulText lexicons present some incoherencies that have obliged us to modify each lexicon to some extent. Examples are epicene nouns and adjectives like Spanish *periodista*, Italian *giornalista* and French *journaliste* that didn't have the same encoding as shown in the figure below.

| | | | |
|---|---|---|---|
| FR | journaliste | = | Ncms-- |
| | journaliste | journaliste | Ncfs-- |
| | journalistes | journaliste | Ncfp-- |
| | journalistes | journaliste | Ncmp-- |
| ES | periodista | periodista | Nc.s- |
| | periodistas | periodista | Nc.p- |
| IT | giornalista | giornalista | Nccs- |
| | giornaliste | giornalista | Ncfp- |
| | giornalisti | giornalista | Ncmp- |

Figure 1: epicene nouns in MulText

To avoid inconsistency, some incoherencies or errors were corrected, as these had a negative effect on the statistics conducted on the lexicon, especially with respect to the inflection paradigms. Some inflected forms, e.g, were missing and produced incomplete paradigms.

As for FreeLing, it is an open-source library providing multilingual NLP services such as lemmatization and PoS tagging. The English dictionary was automatically extracted from WSF with minimal hand post-edition and tends to be a little noisy. The Spanish dictionary is hand coded whereas the Italian dictionary has been extracted from Morph-it! Morphy is freely available software for morphological analysis and PoS tagging for German. Lefff 2.1 is a freely available wide-coverage morphosyntactic and syntactic lexicon for French.

The number of lemma and lexicon entries is given in Table 1. We removed the entries containing proper names to avoid the bias introduced by these types of entries.

### 3.2 Description of the corpora set

For our study we have used parallel corpora. Unfortunately, multilingual parallel corpora are hard to come by. As lexical studies on corpora are always biased by the type of discourse represented in the corpus, we used two different sets: the JRC-Acquis v.3.0 and the aligned bible. The XML-encoded JRC-Acquis is a freely available parallel corpus containing EU documents of mostly legal nature in more than 20 languages. Unfortunately, monolingual documents include sentences or paragraphs in one or more languages that are not always marked up and therefore cannot be removed automatically. This fact drastically decreases coverage performance.

As we can see in Table 2, German is the language with the smallest number of word occurrences (tokens) in each of the corpora and with the highest number of different words (types). As we are working with parallel corpora, this indicates that German uses fewer words to express the same thing, while Spanish and French in the JRC Corpus and French in the bible corpus are the languages with more words. This fact isn't surprising since German has a very productive morphological composition that enables the creation of new words.

A curious fact is that English is the language with a smaller proportion of types, indicating that the vocabulary used is less variable than in other languages. Italian shows a more varied vocabulary, especially in the bible corpus.

## *4 Comparing vocabulary size and coverage*

### 4.1 Comparing vocabulary growth

The vocabulary growth is an indicator of the difficulty to build an appropriate lexicon for a given language. Table 3 gives the number of words (forms) needed in order to have a certain static coverage. Static coverage indicates the percentage of tokens in the corpus mapped by the lexicon, while dynamic coverage refers to the types (Mérialdo, 1988). Cells in grey indicate the largest number of words needed to reach the given coverage while the cells in italics indicate the smallest number of required words.

We can see that German has an extensive vocabulary. English uses a smaller vocabulary when the coverage is higher than 70%. The needed vocabulary can be twice as big in one language as in another.

### 4.2 Comparing coverage

In this section we present the dynamic and static coverage for each lexicon described in 3.1 and run on the corpora mentioned in 3.2. As we can see in Table 4, results on the bible are better than the ones in the JRC corpus, because the JRC corpus represents a quite technical discourse and also because of the noise reported in 3.2.

In German it seems to be more difficult to achieve a good coverage than in other languages. The German MulText and the German Morphy lexicons score a dynamic coverage of 0.35 and 0.59 and a static one of 0.83 and 0.89, while the highest score for dynamic and static coverage is achieved in French for both lexicons (0.82 and 0.83 dynamic coverage and 0.96 and 0.95 static coverage). Italian is the Latin language with the weakest coverage, while the Spanish FreeLing and the French Lefff achieve a dynamic coverage of 0.78 and 0.83; Italian shows 0.70 of dynamic coverage. The static coverage is also lower than for the other Latin languages (0.91 in the bible corpus against 0.94 for Spanish and 0.95 for French). The question arises as to whether this difference is due to the quality of the lexicon or to the language itself.

### 4.3 Comparing statistics on lexicons with the same coverage

In order to compare the lexicon on the same basis, we have extracted new lexicons from the original ones that are needed to cover 60% of the tokens in the JRC corpus and 70% in the bible corpus. These lexicons were generated extracting all the lemmas that could be associated with a given form (in French, *portes* is associated with the noun *porte* and the verb *porter*). Then we derived all the inflectional forms associated with these lemmas. After generating those lexicons, we created automatically the associated lemmas with all their corresponding inflections.

Note that we have limited our study to open classes (without adverbs). Table 5 shows that in German more than double the number of lemmas are needed to achieve the same coverage as for Spanish. For coverage of 70% in the bible corpus in German we needed 482 MulText lemmas and 478 Morphy lemmas whereas in French we only needed 119 MulText lemmas and 202 FreeLing lemmas. At the same time, results indicate that German is the language with the largest number of open class tags (MulText 393, Morphy 175, JRC) and English the one with the smallest amount of tags (36 in MulText, 12 in FreeLing, JRC). Latin languages do not have the same number of tags, but they all have a number greater than for English and smaller than for German. Surprisingly, Italian has many more tags than other Latin languages.

## *5 Analysis of the morphosyntactic complexity*

### 5.1 Comparing morphosyntactic ambiguity

Concerning ambiguity, Table 6 gives the average number of possible tags for a given form using MulText. This is evaluated considering both types and tokens. Moreover, simple PoS, that is A, N or V and complete tags (for instance Ncms-) are considered.

Spanish seems to be the less ambiguous language. The number of tags per form for German is very high due to the choices made when building the original lexicons as explained below (5.2). Italian is shown to be the most ambiguous language regarding the number of PoS. Again the question arises as to whether this result is a consequence of the quality of the lexicons, especially since the Italian FreeLing has been built up automatically.

### 5.2 Comparing inflection paradigms

After generating the lexicons needed for a given coverage (see section 4.3) we generated automatically the number of inflection paradigms and rules to inflect the lemmas. These rules were induced from the obtained lexicon. The idea is to get a lexicon with a lemma and an inflection rule that can be applied to generate a form-lemma-tag lexicon.

Table 7 presents the number of paradigms per language in the bible corpus, the number of paradigms per PoS, the average number of inflections for the total paradigms and the number of inflections per PoS paradigms. We also expose the number of endings per rule that are added to the stem and the number of

endings that are removed in the inflection process.

To our surprise we can see that in Spanish a smaller number of paradigms is needed than for other languages; depending on the corpus and lexicon used, Spanish is equivalent to English. We expected that only English would show a small number of paradigms. Regarding verbal inflection, although Spanish has fewer paradigm rules than English (between 18 and 22 for Spanish for the bible corpus and between 31 and 32 in English), each paradigm generates a high number of forms: while English obtains less than 7 inflected forms per paradigm, Spanish has an average of between 65 (FreeLing) and 157 (MulText)! Again the striking difference between the lexicons can be explained by the fact that the encoding philosophy diverges a lot from one project to another. Whereas FreeLing handles Spanish verbal cliticization with a special module for morphological analysis, Multext includes all the inflected forms with clitic as *bebiéndolo* in the lexicon.

The same can be argued for the differences noted for German MulText and Morphy. In Morphy, e.g, all verbs with a separable particle are lemmatized to the verb without particle, as *zurückgekommen* lemmatized to *kommen* instead of to its infitive *zurückkommen* as encoded in Multext. This explains big differences in verbal paradigms. The number of verbal endings in Morphy, e.g., is multiplied by a factor of 30!

Yet another interesting observation is the number of characters to be deleted in the inflection process. The average number in Spanish is lower than in French in Italian, but English is still the language with the lowest average.

## 6 Conclusions and perspectives

The different figures highlighted in this paper provide a great deal of information on languages and are sometimes quite surprising, as the low number of inflection paradigms needed for Spanish.

But beyond the figures themselves, our results indicate how difficult it is to build up harmonized multilingual lexicons; that is, to create lexicons according to a common tagset (even when language specific attributes are foreseen). Even lexicons developed for the same purposes and under the same project as MulText and FreeLing, do not always fulfil this requirement.

The errors found in the lexicons are another problem for our study. Sometimes they are due to the use of automatic procedures as was the case for the English FreeLing, that was generated using an automatically created lexicon whereas in Spanish these tasks were handled by linguists.

In this paper we present a first approach for the automatic comparison of the difficulties of different languages for NLP applications. Lexicons are indeed of paramount importance for NLP. The development of these resources is a complex task and it is interesting to find clues to help predict the degree of difficulty. The approach presented here makes use of existing lexicons and shows how encoding differences and errors impede the obtention of reliable results. A more challenging method would be to predict the difficulty without previous knowledge. We plan to run further studies using tools to automatically induce morphology from corpora like *Linguistica* (Goldsmith 2006) and to compare the obtained results with the ones presented here.

Moreover, as multilingual parallel corpora are too specific and difficult to come by, further work will be carried out using comparable corpora. Here again, the comparison between the results obtained with parallel and comparable corpora will enable us to determine whether it is possible to evaluate the difficulty of creating morphosyntactic lexicons without previous resources.

## 7 References

Atserias, J., Casas, B., Comelles, E., González, M., Padró, L., Padró, M., 2006. FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. *Proceedings of the fifth international conference on Language Resources and Evaluation* (LREC 2006), ELRA. Genoa, Italy. May, 2006.

Comrie, B., 1989. *Language universals and linguistic typology*. Chicago: The University of Chicago Press.

Goldsmith, J., 2006. An algorithm for the unsupervised learning of morphology. *Natural Language Engineering 12*. 1-19.

Ide, N., Véronis, J., 1994. MULTEXT: Multilingual Text Tools and Corpora. *Proceedings of the 15th International*

*Conference on Computational Linguistics*, COLING'94, Kyoto, Japan, 588-92.

Lezius, W., 2000. Morphy - German Morphology, Part-of-Speech Tagging and Applications in Ulrich Heid; Stefan Evert; Egbert Lehmann and Christian Rohrer, editors, *Proceedings of the 9th EURALEX International Congress* pp. 619-623 Stuttgart, Germany.

Mérialdo, B., 1988. Multilevel decoding for very-large-size-dictionary speech recognition, IBM Journal of Research and Development, v.32 n.2, p.227-237, March 1988.

Pirkola, A., 2001. "Morphological Typology of Languages for IR", *Journal of Documentation*, 57, 2001, 330-348.

Resnik, P., Broman Olsen, M., and Diab, M., 1999. The Bible as a parallel corpus: Annotating the "Book of 2000 Tongues." *Computers and the Humanities* 33, 1–2 (1999) 363–379.

Sagot, B., Clément, L., Villemonte de la Clergerie, E., Boullier, P., 2006. The Lefff 2 syntactic lexicon for French: architecture, acquisition, use. In the *Proceedings of the Language Resources and Evaluation Conference*, LREC'06, Gênes

Schone, P., & Jurafsky, D., 2001. Knowledge-Free Induction of Inflectional Morphologies. *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics* (NAACL-01).

Steinberger, R., Pouliquen, B., Widiger, A., Ignat , C., Erjavec ,T., Tufis, D., Varga, D., 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *Proceedings of the 5th International Conference on Language Resources and Evaluation* (LREC 2006). Genoa, Italy, 24-26 May 2006.

Trost, H., 2003. Computational Morphology. In: Ruslan Mitkov (editor), *The Oxford Handbook of Computational Linguistics*, pp. 25-47. Oxford University Press.

Whaley, L.J., 1997. Introduction to typology: the unity and diversity of language. Thousand Oaks - London - New Delhi: Sage Publications.

## *8 Annex : Tables*

| Lexicon | Language | Lemma | Entries |
|---|---|---|---|
| MulText EN | EN | 14,639 | 66,215 |
| FreeLing EN | EN | 40,219 | 67,213 |
| MulText ES | ES | 18,027 | 510,711 |
| FreeLing ES | ES | 76,201 | 668,816 |
| MulText IT | IT | 10,238 | 232,079 |
| FreeLing IT | IT | 40,277 | 437,399 |
| MulText DE | DE | 12,733 | 233,858 |
| Morphy DE | DE | 91,311 | 4,055,789 |
| MulText FR | FR | 28,627 | 306,795 |
| Lefff FR | FR | 56,917 | 472,582 |

Table 1: Number of lexicon entries

| Corpus | Bible | | | | | JRC | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Language | **de** | **en** | **es** | **fr** | **it** | **de** | **en** | **es** | **fr** | **it** |
| nb of types in corpus | 26,380 | 14,679 | 25,238 | 21,385 | 30,498 | 58,800 | 45,079 | 50,202 | 47,858 | 50,388 |
| nb of tokens in corpus | 649,488 | 816,270 | 841,765 | 929,211 | 855,329 | 1,458,661 | 1,524,011 | 1,634,317 | 1,612,744 | 1,557,464 |

Table 2: Number of words and types in the corpora

| | Bible | | | | | JRC | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Coverage** | **de** | **en** | **es** | **fr** | **it** | **de** | **en** | **es** | **fr** | **it** |
| **60%** | 202 | 106 | 108 | 102 | 186 | 368 | 253 | 217 | 238 | 321 |
| **70%** | 502 | 235 | 264 | 220 | 416 | 792 | 538 | 515 | 527 | 652 |
| **80%** | 1321 | *562* | 742 | 621 | 1117 | 2488 | 1257 | 1307 | 1291 | 1665 |
| **90%** | 4069 | *1606* | 2675 | 2274 | 3822 | 10631 | *5074* | 5738 | 5460 | 6774 |
| **99%** | 20847 | *8672* | 18054 | 13617 | 23230 | 50832 | *36365* | 40499 | 38345 | 41515 |
| **100%** | 26380 | *14679* | 25238 | 21385 | 30498 | 58800 | *45079* | 50202 | 47858 | 50388 |

Table 3: Vocabulary growth according to the static coverage

| Lexicon | MulText | Morphy | MulText | FreeLing | MulText | FreeLing | MulText | Lefff | MulText | FreeLing |
|---|---|---|---|---|---|---|---|---|---|---|
| Language | **de** | **de** | **en** | **en** | **es** | **es** | **fr** | **fr** | **it** | **it** |
| **JRC** | | | | | | | | | | |
| **Known types** | 9190 | 13154 | 9733 | 8862 | 13296 | 15574 | 13768 | 14538 | 11123 | 14285 |
| **Unknown types** | 49533 | 45569 | 35329 | 36199 | 36817 | 34538 | 33971 | 33201 | 39235 | 36073 |
| **Dynamic coverage** | 0.16 | 0.22 | 0.22 | 0.20 | 0.27 | 0.31 | 0.29 | 0.30 | 0.22 | 0.28 |
| **Static coverage** | 0.63 | 0.68 | 0.74 | 0.75 | 0.76 | 0.81 | 0.82 | 0.80 | 0.76 | 0.74 |
| **Bible** | | | | | | | | | | |
| **Known types** | 9111 | 15446 | 9747 | 8498 | 16267 | 19779 | 17515 | 17804 | 13582 | 21287 |
| **Unknown types** | 17238 | 10904 | 4932 | 6180 | 8969 | 5456 | 3824 | 3535 | 16914 | 9209 |
| **Dynamic coverage** | 0.35 | 0.59 | 0.66 | 0.58 | 0.64 | 0.78 | 0.82 | 0.83 | 0.45 | 0.70 |
| **Static coverage** | 0.83 | 0.89 | 0.93 | 0.90 | 0.90 | 0.94 | 0.96 | 0.95 | 0.84 | 0.91 |

Table 4: Coverage of the lexicons

| Lexicon | MulText | Morphy | MulText | FreeLing | MulText | FreeLing | MulText | Lefff | MulText | FreeLing |
|---|---|---|---|---|---|---|---|---|---|---|
| Language | de | de | en | en | es | es | fr | fr | it | it |
| **JRC** | | | | | | | | | | |
| **Lemmas (A, N, V)** | 1533 | 915 | 577 | 524 | 393 | 328 | 397 | 373 | 557 | 748 |
| **Tags (A, N, V)** | 393 | 175 | 36 | 12 | 131 | 194 | 176 | 112 | 123 | 277 |
| **Adjectives** | 484 | 204 | 109 | 126 | 81 | 41 | 82 | 48 | 106 | 165 |
| **Nouns** | 742 | 563 | 303 | 258 | 214 | 209 | 228 | 248 | 302 | 398 |
| **Bible** | | | | | | | | | | |
| **Lemmas (A, N, V)** | 482 | 478 | 229 | 280 | 129 | 205 | 119 | 202 | 510 | 350 |
| **Tags (A, N, V)** | 391 | 173 | 38 | 12 | 131 | 188 | 170 | 112 | 122 | 278 |
| **Adjectives** | 146 | 122 | 32 | 55 | 11 | 17 | 12 | 28 | 85 | 73 |
| **Nouns** | 225 | 253 | 112 | 136 | 80 | 136 | 82 | 138 | 272 | 190 |
| **Verbs** | 111 | 103 | 85 | 89 | 38 | 52 | 25 | 36 | 153 | 87 |

Table 5: Number of lexicon entries for a coverage of 60% for the JRC Corpus and 70% for the bible

| **Bible** | **de** | **en** | **es** | **fr** | **it** |
|---|---|---|---|---|---|
| **Average tags by form in lexicon** | 5,80 | 1,44 | *1,23* | 1,58 | 1,61 |
| **Average tags by form in corpus** | 4,10 | 1,51 | *1,33* | 2,21 | 2,09 |
| **Average PoS by form in lexicon** | 1,32 | 1,29 | *1,09* | 1,11 | 1,26 |
| **Average PoS by form in corpus** | 1,52 | 1,33 | *1,24* | 1,42 | 1,67 |
| **JRC** | **de** | **en** | **es** | **fr** | **it** |
| **Average tags by form in lexicon** | 6,90 | 1,35 | *1,29* | 1,54 | 1,73 |
| **Average tags by form in corpus** | 4,90 | 1,43 | *1,37* | 2,06 | 2,01 |
| **Average PoS by form in lexicon** | 1,28 | 1,26 | *1,15* | *1,15* | 1,37 |
| **Average PoS by form in corpus** | 1,40 | 1,34 | *1,26* | 1,42 | 1,64 |

Table 6: Grammatical ambiguity rates (MulText)

| **Bible** | MulText | Morphy | MulText | FreeLing | MulText | FreeLing | MulText | Lefff | MulText | FreeLing |
|---|---|---|---|---|---|---|---|---|---|---|
| | **de** | **de** | **en** | **en** | **es** | **es** | **fr** | **fr** | **it** | **It** |
| Total paradigms | 292 | 184 | 56 | 50 | 34 | 51 | 40 | 54 | 79 | 61 |
| Paradigms (A) | 119 | 17 | 8 | 7 | 5 | 7 | 6 | 12 | 11 | 9 |
| Paradigms (N) | 65 | 64 | 16 | 12 | 11 | 22 | 15 | 21 | 27 | 22 |
| Paradigms (V) | 108 | 103 | 32 | 31 | 18 | 22 | 19 | 21 | 41 | 30 |
| Inflections per paradigm | 55.97 | 139.04 | 4.66 | 4.98 | 85.03 | 29.47 | 24.93 | 23.33 | 29.51 | 36.21 |
| Inflections per paradigm (A) | 104.22 | 152.12 | 2.75 | 2.00 | 5.20 | 2.14 | 4.67 | 4.75 | 5.73 | 3.56 |
| Inflections per paradigm (N) | 9.14 | 9.30 | 2.12 | 1.92 | 2.91 | 2.45 | 2.73 | 3.05 | 2.59 | 2.00 |
| Inflections per paradigm (V) | 30.98 | 217.50 | 6.41 | 6.84 | 157.39 | 65.18 | 48.84 | 54.24 | 53.61 | 71.10 |
| Endings (A, N, V) | 1161 | 14258 | 54 | 47 | 704 | 416 | 274 | 342 | 370 | 355 |
| Endings (A) | 752 | 102 | 6 | 2 | 4 | 4 | 5 | 13 | 14 | 9 |
| Endings (N) | 50 | 58 | 7 | 2 | 10 | 9 | 8 | 16 | 10 | 7 |
| Endings (V) | 483 | 14110 | 43 | 43 | 698 | 411 | 267 | 327 | 356 | 347 |
| Average endings length (A, N, V) | 4.59 | 8.99 | 1.28 | 1.21 | 6.12 | 3.88 | 3.36 | 3.35 | 2.99 | 3.08 |
| Average endings length (A) | 5.40 | 4.24 | 0.91 | 0.36 | 0.85 | 0.80 | 0.71 | 1.14 | 2.06 | 1.66 |
| Average endings length (N) | 1.40 | 1.54 | 0.85 | 0.22 | 1.28 | 1.07 | 0.95 | 1.16 | 0.64 | 0.43 |
| Average endings length (V) | 2.13 | 9.74 | 1.40 | 1.38 | 6.22 | 4.02 | 3.54 | 3.58 | 3.10 | 3.16 |
| Average deleted char (A, N, V) | -2.33 | -5.05 | -0.46 | -0.50 | -1.66 | -1.78 | -2.41 | -2.53 | -2.14 | -2.30 |
| Average deleted char (A) | -2.54 | -1.07 | -0.23 | -0.21 | -0.35 | -0.13 | -0.04 | -0.54 | -0.70 | -1.41 |
| Average deleted char (N) | -0.46 | -0.56 | -0.29 | -0.04 | -0.31 | -0.43 | -0.29 | -1.48 | -0.54 | -0.39 |
| Average deleted char (V) | -1.89 | -5.63 | -0.51 | -0.57 | -1.69 | -1.84 | -2.57 | -2.68 | -2.23 | -2.35 |

Table 7: Inflection paradigms for a lexicon covering 70% of the bible