On the robust measurement of inflectional diversity

Conference Paper · January 2015

CITATIONS
0

READS
25

Aris Xanthos
University of Lausanne
19 PUBLICATIONS **167** CITATIONS
SEE PROFILE

Guillaume Guex
University of Lausanne
15 PUBLICATIONS **22** CITATIONS
SEE PROFILE

This is a preprint. A revised version has been published as Xanthos, A. & Guex, G. (2015). On the robust measurement of inflectional diversity. In A. Tuzzi, M. Benesova, & J. Macutek (Eds), *Recent Contributions to Quantitative Linguistics* (p. 241-254). Berlin: De Gruyter.

On the robust measurement of inflectional diversity

Aris Xanthos and Guillaume Guex[1]

University of Lausanne

Abstract

Lexical diversity measures are notoriously sensitive to variations of sample size. In order to deal with this issue, most recent approaches involve the computation of the resampled variety of lexical units, i.e. their average variety in random subsamples of fixed size drawn from the corpus. This technique, which has been shown to effectively reduce the influence of sample size variations, has been further applied to measures of inflectional diversity such as the average number of wordforms per lexeme, also known as the mean size of paradigm (MSP) index, thus yielding a so-called *normalized* MSP value.

In this contribution we argue that, while random sampling can indeed be used to increase the robustness of inflectional diversity measures, using a fixed subsample size as in the normalized MSP approach is only justified under the hypothesis that the corpora that we compare have the same degree of lexical diversity–or to be more precise, of lexematic diversity. In the more general case where they may have differing degrees of lexematic diversity, a more sophisticated strategy can and should be adopted.

Based on this reasoning, a novel approach to the measurement of inflectional diversity is proposed, aiming to cope not only with variations of sample size, but also with variations of lexematic diversity. The robustness of this new method is then empirically assessed and compared to that of the standard normalized MSP algorithm, based on text samples generated by a probabilistic model whose degree of lexematic diversity is artificially controlled without altering its degree of inflectional diversity. The results suggest that although there is still room for improvement, the proposed methodology considerably attenuates the impact of lexematic diversity discrepancies on the measurement of inflectional diversity.

Keywords: inflectional diversity, mean size of paradigm, MSP, RMSP, lexical diversity, robustness, random sampling.

1. Introduction

1.1 Lexical diversity, sample size, and random sampling

The measurement of lexical diversity is one of the most studied topics in quantitative linguistics. The basic ingredient of all diversity measures is *variety*, namely the number *V* of

[1] Corresponding author: Aris Xanthos, University of Lausanne, Anthropole, CH–1015 Lausanne, aris.xanthos@unil.ch.

distinct lexical units in a text sample. It is well-known that V is critically dependent on the number N of tokens in the sample, so that samples of differing sizes cannot be directly compared based on this index. Many studies have tried to circumvent this issue using instead the *type-token ratio* TTR := V/N. However, TTR is also dependent on N in a non-linear fashion and the same holds about the various transforms of TTR that have been proposed by Guiraud (1954), Herdan (1960), and several others (see e.g. Tweedie & Baayen, 1998 and references cited therein).

Many recent approaches to diversity measurement rely on a different way of compensating for sample size variations, based on an idea formulated seventy years ago by Johnson (1944): computing and reporting the *average* TTR (or, equivalently, variety) in a number of fixed-size subsamples drawn from the sample under consideration. In Johnson's original proposal (sometimes called *mean segmental* TTR), subsamples are defined as contiguous, non-overlapping sequences of l_{sub} tokens ($1 \leq l_{sub} \leq N$). Consequently, the number n_{sub} of subsamples is determined by the integer division $\lfloor N/l_{sub} \rfloor$. Furthermore, when l_{sub} is not a factor of N, adopting this sampling scheme implies discarding a "residue" of at most $l_{sub}-1$ tokens.

The constraint that subsamples should be made of contiguous tokens has usually been relaxed in later studies, as illustrated by Dubrocard (1988), where the N tokens composing the sample are randomly assigned to the subsamples, regardless of their position in the text. Malvern & Richards (1997) have further advocated a sampling procedure where each subsample is built by drawing tokens without replacement in the text–similarly to Johnson's or Dubrocard's method–but a given token may occur in any number of subsamples (including 0). The consequence of this change in design is that the number n_{sub} of subsamples becomes an actual parameter, whose value may be set to an arbitrary large number, irrespective of subsample size l_{sub}.

Malvern & Richards (1997) proceed with the specification of a sophisticate approach that has become the current *de facto* standard for measuring lexical diversity. This approach, called VOCD, relies on the calculation of the average TTR in subsamples of increasing size (35,36,…,50 tokens), in order to build a so-called "empirical" TTR curve. A curve-fitting procedure is then applied to find the "theoretical" curve which matches the empirical one most closely, among a family of curves generated by the variation of a single parameter in a mathematical model of the relationship between sample size and TTR. The parameter value generating the curve with the best fit is eventually reported as the measured diversity.

The usefulness of the VOCD algorithm has been seriously challenged in a recent contribution by McCarthy & Jarvis (2007). These authors convicingly argue that (i) the curve-fitting procedure underlying VOCD has no other use than smoothing the fluctuations induced by random sampling; and (ii) that a better way of achieving this effect is to calculate analytically the *expected* TTR in *all* possible subsamples of a given size–a calculation whose details (based on the hypergeometric law) have been specified already by Serant (1988), and essentially ignored for the next two decades.

1.2 Inflectional diversity

The notion of *inflectional* diversity relies on the distinction between *inflected wordforms* or simply *forms* (such as *walk*, *walked*, and *walking*) and *lexemes* or *lemmas*, i.e. the abstract lexical categories to which related wordforms belong (such as the verb conventionally referred to using the infinitive TO WALK). In what follows, we will conventionally denote the number of distinct wordforms in a sample by F, and we will call this number the sample's

wordform variety. Similarly, the number of distinct lexemes will be denoted by *L* and called *lexematic* variety. Both quantities capture distinct but interrelated aspects of lexical diversity.

The measurement of inflectional diversity has a much shorter history than that of its lexical counterpart. In particular, many studies have simply used the average number of wordforms per lexeme, also known as the *mean size of paradigm*[2] (see Xanthos & Gillis, 2010 and references cited therein), defined as MSP := *F*/*L*, i.e. the ratio of wordform variety to lexematic variety. However, being a type/type ratio, MSP is easily shown to inherit its components' dependence on sample size. As such, it cannot either be used for directly comparing samples of differing sizes.

To the best of our knowledge, there have been only two proposals for the measurement of inflectional diversity that explicitly take into account the issue of dependence to sample size. The first is based on VOCD (see section 1.1) and due to Malvern, Richards, Chipere & Durán (2004). Based on the observation that VOCD consistently returns slightly lesser values when applied to lexemes than to wordforms, Malvern and colleagues propose to use the *difference* between these two indices as a measure of inflectional diversity (which they call ID). Xanthos & Gillis (2010) have argued that in spite of its promises, this measure suffers from several shortcomings, chief among which are that "the unit in which ID is expressed has no meaningful interpretation" (p.179) and that:

> in the context of an increase in lexical diversity..., ID is liable to detect spurious increases in inflectional diversity–increases that are mere side-effects of the subtractive definition of the measure (p.180).

On these grounds, Xanthos & Gillis have put forward an alternate measure which is easier to compute and, arguably, to interpret. Building on the idea of using random sampling to deal with the dependence on sample size, they define the *normalized* MSP as the average MSP computed in n_{sub} subsamples of l_{sub} tokens drawn randomly from the original sample.[3] They provide empirical evidence showing that using random sampling significantly increases the measurement's robustness with regard to variations of sample size, while preserving its ability to detect variations of inflectional diversity.

It should be noted that as far as we know, the problem of analytically calculating the *expected* MSP in *all* possible subsamples of a given size has not yet been solved. Our own preliminary investigations have given us no reason to believe that it has a solution as simple and elegant as what Serant (1988) has offered for lexical variety.

1.3 Normalized MSP and lexematic diversity

While normalized MSP, as defined above, appears to be robust with regard to sample size variations, the same does not hold for variations of lexical diversity. Xanthos & Gillis (2010) briefly touch upon the issue of the relation between normalized MSP and lexematic diversity:

> given that sample size remains constant, any increase in the diversity of lemmas is matched by a corresponding decrease in the average frequency of lemmas. As more distinct lemmas occur, each of them has less frequent occurrences, which means less space for deploying the variety of its inflected wordforms. Rarer inflections are thus less likely to appear in the sample, and on average a lemma will tend to have a smaller number of distinct wordforms. Overall, a *decrease* in inflectional diversity should occur as a result of the increase in lexical diversity. (p.179).

---

[2] A lexeme's *paradigm* is the set of wordforms belonging to this lexeme.

[3] The same sampling scheme as Malvern & Richards (1997) is used (cf. section 1.1).

In the present contribution, we wish to take this line of reasoning one step further and argue that a sound measure of inflectional diversity should not only be robust with regard to variations of sample size but also with regard to variations of lexematic diversity. Indeed, if normalized MSP reports spurious decreases in inflectional diversity when lexematic diversity increases, it does not fare any better than ID and its own spurious increases (cf. section 1.2).

The first contribution of this study is to introduce an algorithm for computing MSP in such fashion that variations in *both* sample size *and* degree of lexical diversity are being taken into account and compensated for; we optimistically propose to call the resulting measure of inflectional diversity *robust* MSP, or *RMSP*. Secondly, we offer an empirical assessment of the extent to which this new index is less dependent on variations of lexematic diversity than standard normalized MSP (which will henceforth be abbreviated as NMSP); to that effect, we describe a presumably novel method for generating artificial text samples using a probabilistic model whose degree of lexematic diversity can be controlled without modifying its degree of inflectional diversity.

The remainder of this contribution is organized as follows. The next section begins with the justification and specification of the algorithm used for computing the new RMSP index. Then we describe the method that we have designed for controlling the degree of lexematic diversity of artificially generated text samples. We proceed with the description of our experimental setup, including the source data used for our experiments and the way in which they are preprocessed. In section 3, we show the results obtained by NMSP and RMSP, focusing in particular on their relative dependence on lexematic variety. These results are then discussed in section 4 and our main findings briefly summarized in section 5.

2. Method

2.1 The RMSP algorithm

The normalized MSP (NMSP) algorithm attempts to compensate for the dependence of MSP on sample size. It takes as input a set of text samples and computes for each sample the average MSP on n_{sub} subsamples of size l_{sub}. The main constraint is that l_{sub} must be set to a fixed value lesser than or equal to the size l of the smallest sample in the dataset (Xanthos & Gillis, 2010). Normalized versions of lexematic (or wordform) variety (or TTR) can be calculated in the same way, which will be exploited shortly for computing the robust MSP (RMSP) index.

The RMSP algorithm can be thought of as a variant of NMSP where a second layer of normalization is added, in order to compensate not only for the dependence of MSP on sample size, but also on lexematic diversity. Indeed, as noted in section 1.3 above, setting the size of subsamples to a fixed value leads to an underestimation of MSP in samples that have a greater degree of lexematic diversity: in these samples, each lexeme type will have less occurrences on average, which in turn means that it will tend to have less distinct inflected forms–a faithful scale model of the dependency of variety on sample size.

The basic idea underlying the RMSP algorithm is to counterbalance this underestimation issue by adjusting the subsample size l_{sub} separately for each sample, in such fashion that samples with a smaller degree of lexematic diversity (relatively to other samples in the dataset) are assigned a smaller subsample size. In particular, the algorithm attempts to find, for each sample, the subsample size that ensures that lexemes have the same number of tokens on average in all subsamples of all samples; in other words, it seeks to minimize the variance of average lexeme frequency or, equivalently, of its reciprocal, lexematic TTR.

To that effect, a *maximal* subsample size l_{max} is first chosen, with the constraint that it must be lesser than or equal to the size l of the smallest sample in the set. Then the normalized lexematic TTR (henceforth NLTTR) of each sample is computed with a fixed subsample size of l_{max} tokens. The maximal NLTTR value obtained this way determines the target value ($NLTTR_{target}$) that the algorithm consequently tries to reach for each (other) sample in the dataset. In particular, for each sample, the algorithm searches for the subsample size $2 \leq l_{sub} \leq l_{max}$ that is optimal in the sense that the resulting NLTTR value is as close as possible to $NLTTR_{target}$; finally, the NMSP of this sample is computed with the optimal subsample size l_{sub} that has just been found, and the result is reported as the value of the RMSP index for this sample. The algorithm can be described more formally as on Figure 1 below.

RMSP algorithm

Input:

- set S of text samples with size at least l
- maximum subsample size $l_{max} \leq l$

Output: RMSP(s, S, l_{max}) value for each sample $s \in S$

- $NLTTR_{target} \leftarrow \max_{s \in S}(NLTTR(s, l_{max}))$
- for each $s \in S$ do:
  - $l_{low} \leftarrow 2$, $l_{high} \leftarrow l_{max}$
  - $l_{sub} \leftarrow l_{high}$
  - while $NLTTR(s, l_{sub}) \neq NLTTR_{target}$ and $l_{low} \neq l_{high}$ do:
    - $l_{sub} \leftarrow integer((l_{max} + l_{min}) / 2)$
    - if $NLTTR(s, l_{sub}) < NLTTR_{target}$, set l_{high} to l_{sub}
    - else if $NLTTR(s, l_{sub}) > NLTTR_{target}$, set l_{low} to l_{sub}
  - $RMSP(s, S, l_{max}) \leftarrow NMSP(s, l_{sub})$

Figure 1. Algorithm for robust MSP (RMSP) computation.

The following difference between NSMP and RMSP should be stressed. The NMSP value computed for a given sample depends only on the chosen subsample size l_{sub}, so that it can be directly compared with any other NMSP value obtained with the same subsample size. By contrast, the RMSP value of a sample depends not only on the maximum subsample size l_{max} but also on the set of samples with which this sample is compared–or to be precise, on the maximal NLTTR value obtained with a sample of this set for subsample size l_{max} ($NLTTR_{target}$). Consequently, in order to compare this RMSP value with that of a new sample, the following conditions must be met: (i) the new sample must be of size at least l_{max} and (ii) its NLTTR for subsample size l_{max} must be at most $NLTTR_{target}$; if so, the new sample can be processed separately in the same way as each sample of the original dataset. Otherwise, the algorithm must be run again on the entire dataset consisting of the new sample and the old one(s) with which it should be compared.

2.2 Sample generation

In order to evaluate the gain in robustness brought about by the RMSP algorithm, we have designed a method for generating artificial text samples whose degree of lexematic diversity can be controlled without altering their degree of inflectional diversity. This method relies on an $L \times F$ contingency table, where each row corresponds to a lexeme type, each column

corresponds to a wordform type, and each cell gives the count of a pair (*lexeme*, *wordform*).[4] Normalizing over the table's grand total yields a joint probability model that can be used to generate a text sample of size l by drawing l pairs (*lexeme*, *wordform*) with replacement. In what follows, it will be useful to refer to L, F, and F/L as the model's *theoretical* lexematic variety, wordform variety, and MSP respectively.

The models' theoretical lexematic variety can be reduced by aggregating two lexeme types (rows) in the contingency table. Let f and g be the wordform frequency distribution of any two lexemes, ordered by decreasing frequency. By placing the additional constraint that f and g be proportional, we ensure that the aggregated lexeme, defined as the vector sum of f and g, is also proportional to f and g.

In order to substantially decrease the lexematic variety L of the model, we perform $n_{agg} > 1$ aggregations at a time. Now, given that L will be reduced by n_{agg} after n_{agg} aggregations, in order for the theoretical MSP to remain constant, the theoretical wordform variety F should be decreased by $n_{agg} \cdot \text{MSP} = n_{agg}\,(\text{MSP} - 1) + n_{agg}$. The first wordform type of all aggregated lexeme types will contribute to the reduction of F by n_{agg}, so the number of wordform types minus 1 in the aggregated lexeme types should be $n_{agg}\,(\text{MSP} - 1)$. This can be achieved as follows: first, randomly pick lexeme types among those that have more than one wordform type[5], until the wordform "surplus" (i.e. the number of wordform types in the selected lexemes minus the number of selected lexemes) reaches $n_{agg}\,(\text{MSP} - 1)$; then, complete the n_{agg} aggregations by randomly selecting lexeme types among those that have only one wordform type.

We call the process of doing n_{agg} lexeme aggregations as described above an *aggregation round*. After an aggregation round, the modified contingency table can be normalized to build a new joint probability model, which in turn can be used to generate new samples. The process can be repeated as long as there remain enough lexeme types with proportional wordform distribution to aggregate.

2.3 Experimental design

As described in the previous section, our empirical assessment of NMSP and RMSP is based on a probabilistic mechanism for sample generation. The parameters of this mechanism could in principle be themselves generated according to some theoretical model. However, we have rather chosen to estimate them on the basis of natural language data, in order to preserve some degree of resemblance between our experimental design and the "naturalistic" conditions in which the measurement of inflectional diversity is likely to take place.

The data in question are taken from the Project Gutenberg eBook of Eduard Bernstein's *Sozialismus einst und jetzt* (2008). A German text was chosen on the grounds that its degree of inflectional diversity would in principle be relatively high (at least when compared to English, whose inflection is quite limited) so that there would actually be something to measure for our indices. For the same reason, we decided to focus exclusively on the subsystem of *verb* inflection in this corpus.

Bernstein's text was automatically tokenized, lemmatized, and annotated with part-of-speech (POS) tags using TreeTagger (Schmid, 2004). Orange Textable (Xanthos, 2014) was then used to parse the output of TreeTagger and discard all tokens but verbs. The result is a

[4] In practice, these counts are typically derived from an existing text, as described in the next section.
[5] subject to the proportionality constraint discussed above.

list of N = 8106 verb tokens corresponding to F = 2012 wordform types and L = 1078 lexeme types, hence a (raw) MSP of 1.87 forms per lexeme.[6]

Five rounds of 50 lexeme aggregations were made, preserving the theoretical MSP. At each step (starting with no aggregation), 100 text samples of size 500, 1000, 1500, 2000, and 2500 were produced, for a total of 3000 text samples. NMSP was computed with n_{sub} = 1000 subsamples of size l_{sub} = 100, 200, 300, and 400. RMSP was computed with n_{sub} = 1000 subsamples and maximum size l_{max} = 100, 200, 300, and 400.

3. Results

As shown on Figure 2, while lexeme aggregation reduces the model's theoretical lexematic and wordform variety by more than 20% (from 1078 to 828 lexeme types and from 2012 to 1540 wordform types), it causes only a slight decrease in theoretical MSP (from 1.866 to 1.860, i.e. less than 0.5%).

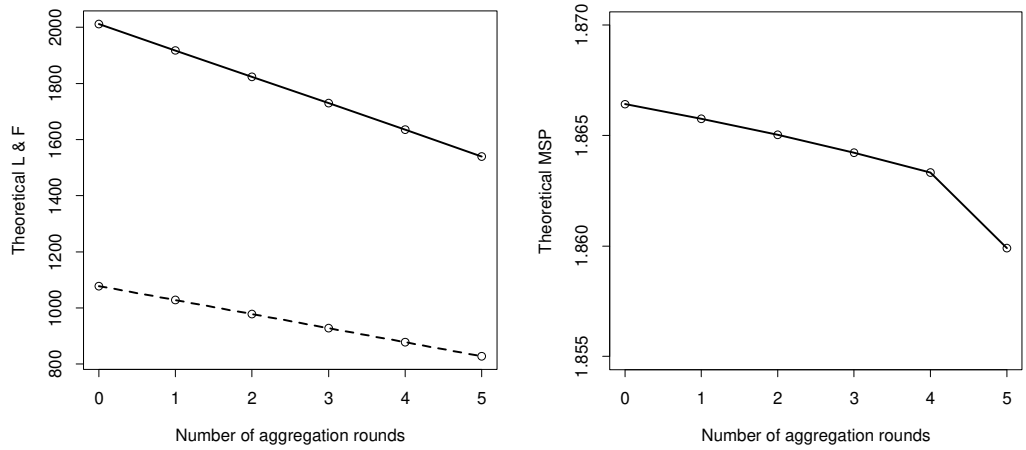

Figure 2. Left: Theoretical values of lexeme (dashed) and wordforms (solid) types vs aggregation rounds. Right: Theoretical MSP vs aggregation rounds.

Figure 3 confirms the impact of sample size on the lexematic and inflectional diversity of generated samples, as measured by their *raw* (i.e. not normalized) lexematic TTR and MSP. More importantly, the figure shows that lexeme aggregation influences both the lexematic TTR and the MSP of generated samples. In particular, the latter increases as the former decreases, in particular for larger sample sizes: the MSP increase ranges between 3% for samples of 500 tokens and 7.6% for samples of 2500 tokens. One should not be surprised that the raw MSP increases with aggregation rounds although the theoretical MSP remains approximately constant; indeed, the predicted effect of lexeme aggregation on the average MSP of samples of fixed size is exactly the same as the predicted effect of lexeme aggregation on NMSP for a given subsample size.

The normalization performed by the NMSP and RMSP algorithms effectively lessens the dependence of diversity measurement on sample size, as indicated by the overlap of curves on Figure 4 (obtained with l_{sub}, l_{max} = 100). The figure also shows that the reported RMSP is

[6] Note that homophonous wordforms belonging to different lexemes are treated as distinct wordform types (e.g. *gehabt*, which can be the past participle of HABEN 'to have' or GEHABEN 'to behave').

systematically lower than the corresponding NMSP. Finally, it can be seen that both measures are affected by lexeme aggregation, although not to the same extent.

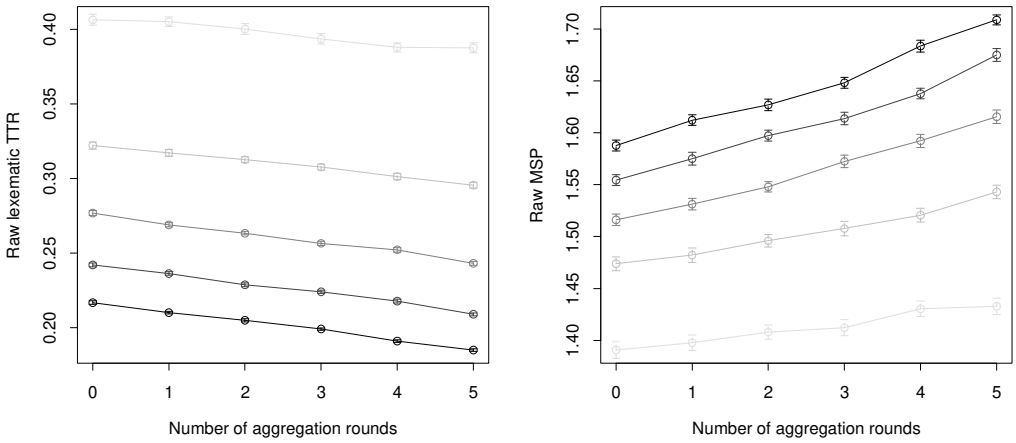

Figure 3. Left: Raw lexematic TTR vs aggregation rounds. Right: Raw MSP vs aggregation rounds. On both figures, light to dark represents samples from size 500 to 2500.

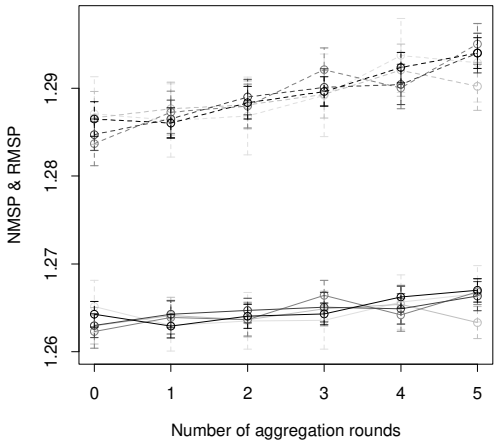

Figure 4. NMSP (dashed) and RMSP (solid) vs aggregation rounds (l_{sub}, l_{max} = 100). Light to dark represents samples from size 500 to 2500.

Figure 5 shows the behavior of NMSP and RMSP for l_{sub}, l_{max} = 100, 200, 300, and 400 tokens (aggregating the results observed for all sample sizes). While both measures are increasing with lexeme aggregations for all values of l_{sub} and l_{max}, the increase is consistently lesser for RMSP than for NMSP. The visual impression is confirmed by the results of a Spearman's correlation test assessing the degree of dependence of NMSP and RMSP on the number of aggregation rounds. With the exception of RMSP with l_{max} = 100, both diversity measures always have a significant correlation with the number of aggregation rounds (cf. Table 1). However, the correlation itself is consistently lesser for RMSP.

4. Discussion and conclusion

In this contribution, we have argued that while the resampling scheme underlying the normalized MSP (NMSP) measure of inflectional diversity proposed by Xanthos & Gillis (2010) effectively reduces the dependence of the measure on sample size, a more

sophisticated approach is needed when dealing with data samples whose degree of lexematic diversity is heterogeneous. We have introduced a novel algorithm called robust MSP (RMSP), which relies on the idea that what should be normalized is not merely the number of tokens per subsample, but the number of tokens per lexeme in subsamples. To that effect, rather than setting a fixed subsample size for all samples in the considered dataset, the RMSP approach sets the size of subsamples separately for each sample, in such fashion that the variance of average lexeme frequency over all subsamples is minimized.

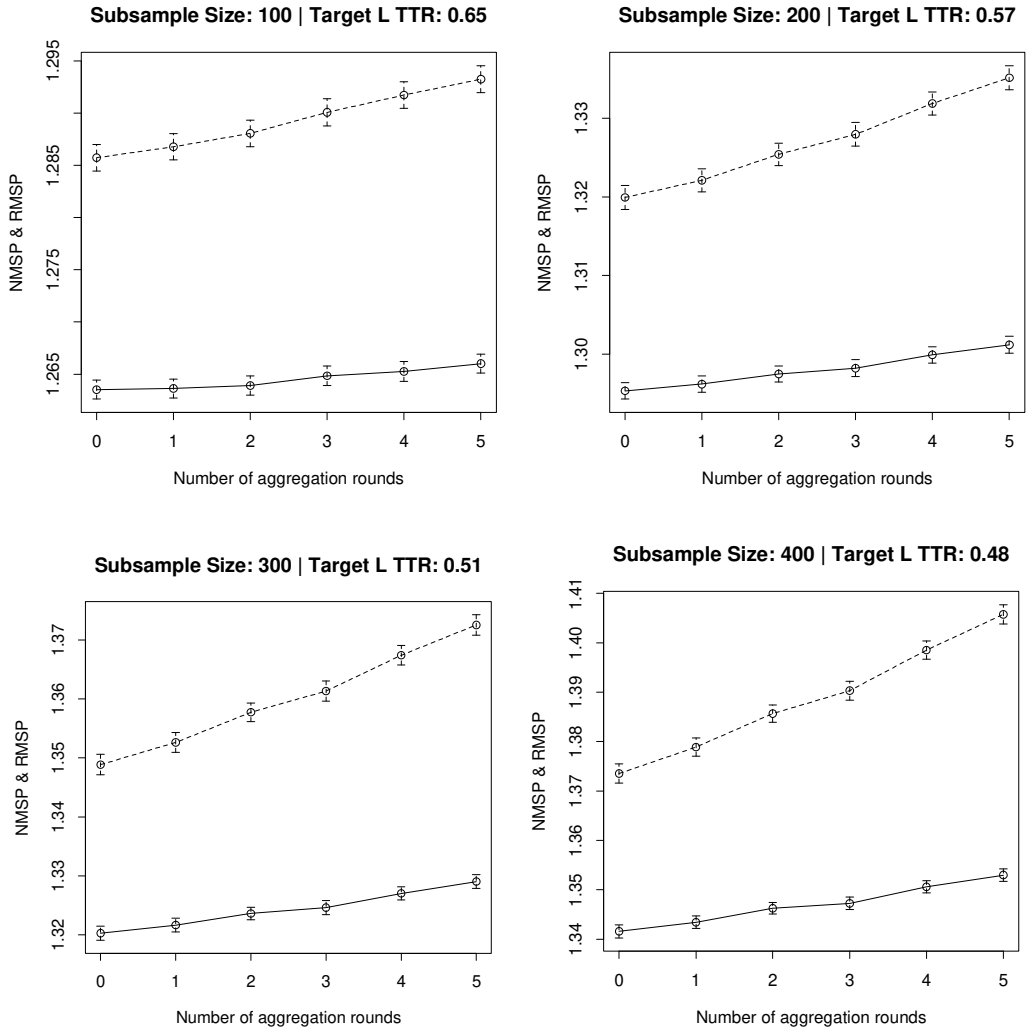

Figure 5. NMSP (dashed) and RMSP (solid) vs aggregation rounds for different subsample size. Results for different sample lengths have been aggregated.

Subsample size	NMSP	RMSP
100	0.190 (p≈0)	0.090 (p≈9e-7)
200	0.318 (p≈0)	0.179 (p≈0)
300	0.414 (p≈0)	0.241 (p≈0)
400	0.484 (p≈0)	0.285 (p≈0)

Table 1. Spearman's correlation between aggregation rounds and NMSP/RMSP.

In order to evaluate the gain in robustness brought about by the RMSP algorithm, we have developed a method for generating artificial text samples (based on lexeme and wordform

frequencies observed in a real text) whose degree of lexematic diversity can be controlled without altering their degree of inflectional diversity. These data have enabled us to show that raw MSP is not only dependent on sample size, but also on variations of lexematic diversity. Applying the NMSP algorithm to the generated samples confirms that while it is much less dependent on sample size than raw MSP, it is also affected by variations of lexematic diversity. Finally, although RMSP is also dependent on lexematic diversity, it proves more robust than NMSP with regard to lexematic diversity fluctuations.

When the samples under consideration are homogeneous from the point of view of their degree of lexematic diversity, RMSP essentially reduces to NMSP (with a slight computational overhead). Otherwise, the RMSP algorithm attempts to compensate for lexematic diversity fluctuations by discarding (through resampling) even more tokens than the standard NMSP algorithm. All other things being equal, discarding more tokens means discarding more types, which explains why the reported values of RMSP are typically lower than those of NMSP. Thus, while RMSP is in principle more widely applicable than NMSP (since it can handle data that display variations of lexematic diversity), it also gets closer to the extreme and absurd case where diversity is evaluated on the basis of a single token. A priority for future research will be to determine the conditions under which the RMSP approach might lead to an information loss so severe that it ultimately fails to provide a meaningful evaluation of inflectional diversity.

References

Bernstein, E. (2008). *Der Sozialismus einst und jetzt. Streitfragen des Sozialismus in Vergangenheit und Gegenwart*. Project Gutenberg. Ed. N. H. Langkau & I. Knoll. Retrieved Dec. 4, 2011 from http://www. gutenberg.org/files/24523/24523-8.txt.

Dubrocard, M. (1988). Evaluation de l'étendue du lexique. Quelques essais de simulation. In P. Thoiron, D. Labbe, & D. Serant (eds.). *Etudes sur la richesse et la structure lexicale. Vocabulary structure and lexical richness.* Paris-Genève: Champion-Slatkine, pp. 43–66.

Guiraud, H. (1954). *Les Caractères Statistiques du Vocabulaire*. Paris: Presses Universitaires de France.

Herdan, G. (1960). *Type-Token Mathematics: A Handbook of Mathematical Linguistics*. The Hague: Mouton & Co.

Johnson, W. (1944). *Studies in language behaviour: I. A program approach*. Psychological Monographs, 56, pp. 1–15.

Malvern, D. & Richards, B. (1997). A new measure of lexical diversity. In A. Ryan & A. Wray (eds.), *Evolving models of language*. Clevedon, UK: Multilingual Matters. pp. 58–71.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Basingstoke: Palgrave MacMillan.

McCarthy, P.M. & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing, 24(4)*: 459–-488.

Serant, D. (1988). A propos des modèles de raccourcissements de textes. In P. Thoiron, D. Labbe, & D. Serant (eds.). *Etudes sur la richesse et la structure lexicale. Vocabulary structure and lexical richness*. Paris-Genève: Champion-Slatkine, pp. 43–66.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Tweedie, F.J. & Baayen, R.H. (1998). How Variable May a Constant be? Measures of Lexical Richness in Perspective. *Computers and the Humanities*, 32(5): 323–352.

Xanthos, A. (2014). Textable: programmation visuelle pour l'analyse de données textuelles. In *Actes des 12èmes Journées internationales d'analy-se statistique des données textuelles (JADT 2014)*, pp. 691-703.

Xanthos, A. & Gillis, S. (2010). Quantifying the development of inflectional diversity. *First Language*, 30(2): 175–198.