A Vector Space Model for Syntactic Distances Between Dialects

Emanuele Di Buccio, Giorgio Maria Di Nunzio, Gianmaria Silvello
Department of Information Engineering, University of Padua
{dibuccio, dinunzio, silvello}@dei.unipd.it

Abstract

Syntactic comparison across languages is essential in the research field of linguistics, e.g. when investigating the relationship among closely related languages. In IR and NLP, the syntactic information is used to understand the meaning of word occurrences according to the context in which their appear. In this paper, we discuss a mathematical framework to compute the distance between languages based on the data available in current state-of-the-art linguistic databases. This framework is inspired by approaches presented in IR and NLP.

Keywords: Digital Geolinguistics, Language Distance, Vector Space Model

1. Motivation and Background

Syntactic comparison across languages is essential in the research field of linguistics. In fact, the study of closely-related varieties has proven to be extremely useful in finding relations between cross-linguistic syntactic differences that might otherwise appear unrelated, and in analysing the linguistic structures in the task of historical reconstruction (Nerbonne and Wiersma, 2006; Colonna et al., 2010). More precisely, syntactic variation studies the ways in which linguistic elements, i.e. words and clitics, are put together to form constituents, that are phrases or clauses. In this context, the analyses of dialectal variation patterns may result in more fine-grained linguistic theories, and empirical dialect data may also help improve the validation process of linguistic theories. Therefore, dialectal variation research may contribute to a better understanding of the inner workings of the human language system (Spruit, 2008). Different dialectal variants do not occur randomly on the territory and geographical patterns of variation are recognizable for an individual syntactic form. In other words, the geographical distribution of an individual syntactic phenomenon is often geographically coherent to a certain extent. This indicates that there might be a relationship between syntactic variation and geographical distance. However, when several distribution patterns of syntactic phenomena are combined for joint analysis, the interpretation of geographical distributions is less clear (Spruit, 2008).

In literature, several approaches for measuring the degree of syntactic differences between varieties have been proposed. The techniques are quantitative by nature, which means that the linguistic data are represented and compared numerically using a function which measures the distance between two points (the varieties). Many of the works in this research field use the Hamming distance (Hamming, 1950) to measure the differences between two or more varieties (Nerbonne and Wiersma, 2006; Spruit, 2008; Spruit, 2006; Spruit et al., 2009). The Hamming distance is calculated between each pair of dialects to obtain a measurement based on binary comparisons between feature variants: the distance is increased by 1 for each feature that is observed in one dialect but not in the other. In (Nerbonne and Wiersma, 2006), instead of binary features, the authors use frequency profiles of trigrams of part-of-speech (POS) categories as indicators of syntactic differences. Nevertheless, since the number of features can be very high, a reduction of the space is usually performed by means of Multidimensional scaling (MDS). MDS is applied to analyse the dialect relationships in the distance matrix. The goal of this procedure in this context is to optimally represent the most differentiating feature variants for each dialect in relation to all other dialects. The results of this reduction to a visible space (two- or three-dimensional space) are visualised with dialect colour maps (Spruit, 2008). For example, in (Spruit et al., 2009), each dialect's distance relationships to all other dialects are reduced to coordinates in a three-dimensional space using the three most important dimensions arising from the MDS analysis. These coordinates optimally represent the original dialect distance relationships. However, they do not directly correspond to actual dialect distances anymore.

More recent approaches try to identify correspondences between languages which are significant against chance and thus call for historical explanation. The computation of the probability of 'mutation' of one language into another is based on the application of genetic algorithms. In genetic algorithms, the basic idea is to cluster the population into a number of groups, based on their similarity with respect to a distance metric (Nguyen et al., 2012). A similar approach is discussed in (Colonna et al., 2010), where the Parametric Comparison Method (PCM) is presented. PCM is a new method of language comparison based on the idea that the core grammar of any natural language can in principle be represented by a string of binary symbols, each symbol coding the value of a linguistic parameter. Such strings of symbols can be unambiguously collated and language distances and chance probability of agreements precisely measured. This approach starts by computing the distance between two varieties as a Jaccard distance (Jaccard, 1901), then, to graphically represent the genetic similarities between populations, MDS is used to project distance matrices in a bi-dimensional space so that the distances between the points approximate the respective degree of dissimilarity

2. A Vector Space for Languages

Following the work of (Spruit, 2006), the term variable (tag) is central to this work. Generally speaking, a variable may be defined as a linguistic unit in which two language varieties can vary. We define a syntactic variable as a form or word order in a syntactic context where two di-

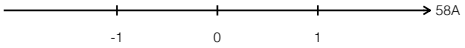

Figure 1: One-dimensional representation of a variety.

alects may differ. Several types of variables can be distinguished; for instance, they can be distinguished according to the linguistic unit to which they refer. The Synctactic Atlas of Italy (ASIt) (Agosti et al., 2010) tag set was defined to support the study on Italian dialects; it includes two different types of tags to capture word-level and sentence-level phenomena. Another example is the set of 192 features made available by The World Atlas of Language Structures (WALS)[1] in which each feature describes one aspect of cross-linguistic diversity (Dryer and Haspelmath, 2013). The basic rationale underlying our approach is to represent language varieties as subspaces. In particular, in this work we exploit one-dimensional subspaces, i.e. vectors, thus obtaining a vector-based representation that shares the same intuition of that proposed in (Salton et al., 1975) to obtain a computational model for Information Retrieval (IR). In IR documents and queries are represented as vectors, whereas here each language variety is represented as a vector $\mathbf{v} \in \mathbb{R}^{|\mathcal{F}|}$ where $\mathcal{F}$ is the set of features (tags) adopted to capture linguistic variations. As an example, let us consider the WALS feature set and focus on a single feature, 58A which captures the 'obligatory possessive inflection' (Bickel and Nichols, 2013). Feature 58A is a binary feature: it can be either present or absent. If we consider only this feature, all the varieties are represented as a vector $v \in \mathbb{R}$, i.e. as a point on the real line; $v = 1$ denotes the case where feature 58A exists, and $v = -1$ the case where the feature is absent. Therefore, all the varieties are represented by one of the two points $\{-1, 1\}$ – see Figure 1. In Figure 2, we can see the geographic distribution of feature 58A.

Let us now consider an additional feature, 107A (Siewierska, 2013), which captures the presence of passive constructions; this is another binary feature – i.e. -1 denote the absence and 1 the presence of this phenomenon. In Figure 3, the map shows a nice geographic distribution of the binary feature 107A; unfortunately, it is not possible to combine features to show the presence and absence of multiple features on the same map. Nevertheless, from a mathematical point of view, each variety can now be represented as a vector $\mathbf{v} \in \mathbb{R}^2$. Since, in this case, both the features are binary, each vector can be in one the four "positions" depicted in Figure 4.

Given these representations, *how can we use them to identify possibly related varieties?* A possible approach is to use the angle θ between the vectors to compute the varieties similarity, or equivalently the cosine of the angle: the assumption is that varieties whose vector representations are close to each other, are related. For instance, if $\mathbf{v}_{11} = (1, 1)$ represents the case where both the phenomena are present (58A=1 and 107A=1), $\mathbf{v}_{00} = (-1, -1)$ the case where both the phenomena are not present, and θ is

[1] http://wals.info

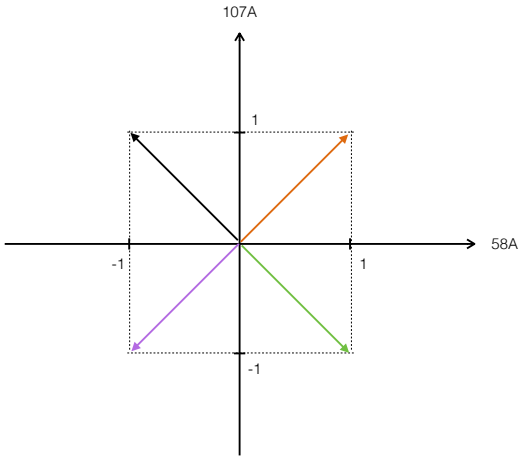

Figure 4: Two-dimensional representation of a variety.

		107A	
		Present (162)	Absent (211)
58A	Exists (43)	14	18
	Absent (201)	72	72

Figure 5: Number of varieties for each 107A - 58A values combination.

the angle between $\mathbf{v}_{11}$ and $\mathbf{v}_{00}$, then $cos(\theta)$=-1. This is, for instance, the case of *Limbu* and *Yoruba* that are spoken respectively in Nepal and Nigeria. The highest similarity value is obtained when comparing varieties with the same vector representation, e.g. all the languages represented by the vector $\mathbf{v}_{11}$ — in this case the $cos(\theta) = 1$.

The above example is based on a very simple representation involving only binary features, while some features can assume more than two values — e.g. feature 17A (Rhythm Types) in WALS has five diverse values. Indeed, this approach is more general because it has no restriction on the values that each feature can take. For example, we may think to integrate into the vector space information about stresses and accents measured in terms of a (continuous) amount of air pressure which gives information about the volume of the voice on that particular sentence.

Moreover, the above representation exploit only the canonical basis: the jth feature is represented by the vector $\mathbf{e}_j$ where $e_{jj} = 1$ and $e_{ij} = 0\ \forall i \neq j$. Similarly to what has been proposed in the context of IR for information object representation (Melucci, 2008), vector space basis other than the canonical basis can be adopted to represent linguistic phenomena in varieties. Indeed, a different vector space basis can be adopted, where features are not considered as independent but their relationships are explicitly modeled – e.g. a basis vector $\mathbf{u} = (1/2, -1/3)$ considers a new feature that represent a specific relationship between feature 58A and feature 107A. This is crucial since, when investigating relationship among closely related varieties, one of the research hypothesis is that the relationship among those varieties can be described by combination of linguis-

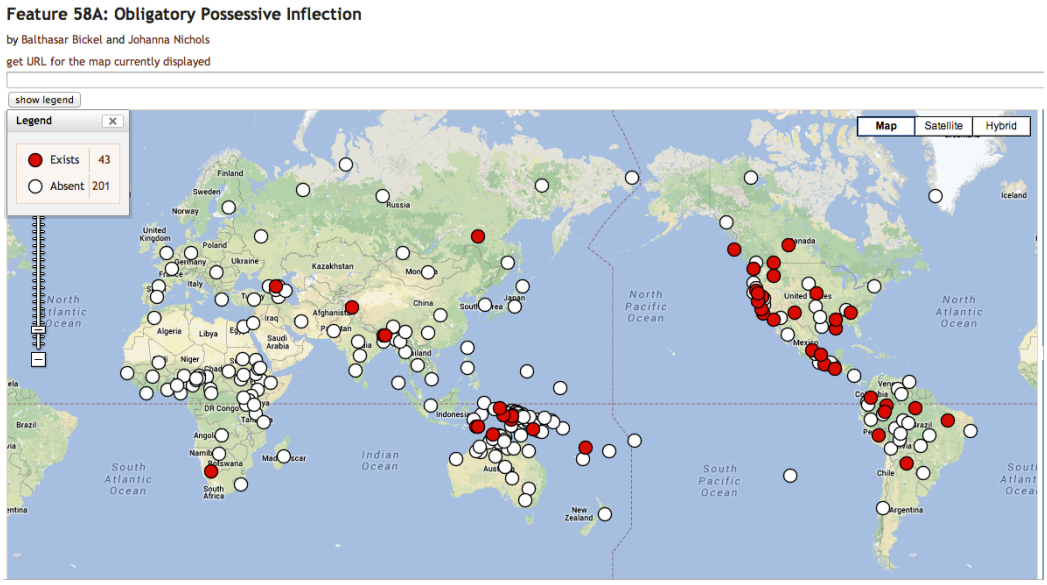

Figure 2: Geographic distribution of feature 58A. WALS database.

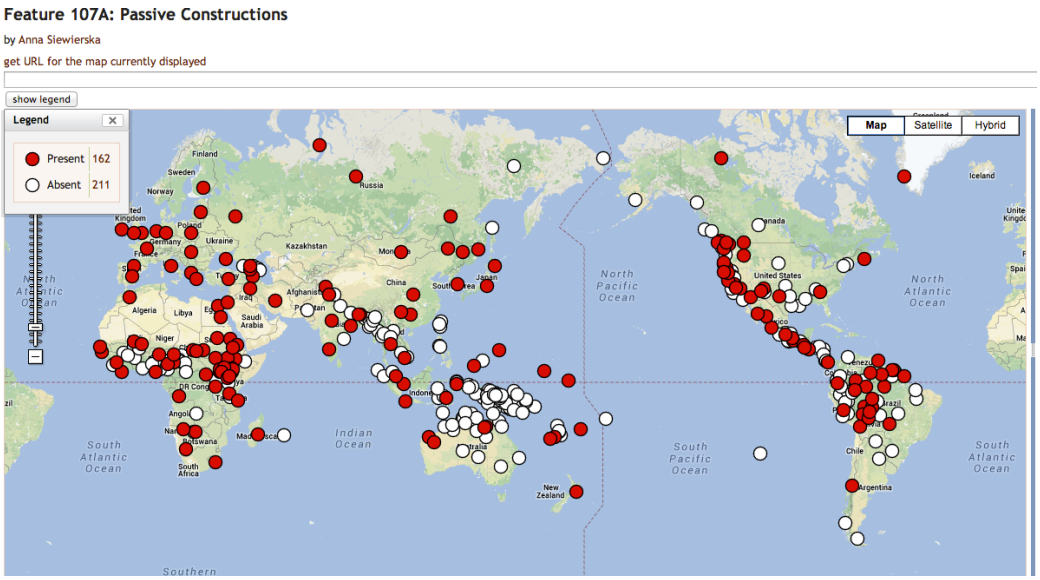

Figure 3: Geographic distribution of feature 107A. WALS database.

tic phenomena. Therefore, a vector space-based representation can help us to model and exploit feature relationships for investigating variety relationships.

Finally, *projection* can be adopted to support the study of specific linguistic phenomena. Indeed, given a vector space basis whose vectors summarize the phenomena of interest in the research hypothesis under investigation, the varieties vectors can be projected onto the subspaces spanned by those basis vectors in order to study relationship among varieties in the context of the considered phenomena.

3. Current and Future Work

In this paper, we presented a brief overview of the current approaches in syntactic comparison for measuring the degree of syntactic differences between languages. The majority of approaches defines a distance based on binary linguistic features (a feature is either present or not in a language), then the matrix of distances between dialects is reduced to a two- or three dimensional space by means of MDS techniques.

Even though the importance of the visualisation on a three-dimensional space is undisputed, we believe that a multi-dimensional space is a much powerful mathematical representation to study fine-grained dialectal differences. The direction we pursue in this paper is at the exact opposite to the one presented in Sec. 1.: we want to build a high multidimensional space by composing small vector spaces. This idea is built on the concept of "clitic clusters" which happens when more than one clitic shows up within a single clause. One very interesting fact about clitic clusters is that the order in which they are in a cluster appears to be random; that is, it is not normally the same order as the corresponding order of full noun phrases, and there is what appears to be random variation between languages as to which ordering restrictions they impose. For example, a third per-

son dative clitic must follow a third person accusative clitic in French, whereas the order must be the other way around in Italian, Spanish and Romanian. [2] For example, the sentence "Martine sends it to him" is translated in:

- Martine le lui envoie (French) (accusative-dative)
- Martina glielo spedisce (Italian) (dative-accusative)
- Martina i-l trimite (Romanian) (dative-accusative)

A first person dative clitic, however, must precede a third person accusative clitic in French (as in the other Romance languages). For example, "Martine sends it to me" becomes:

- Martine me lenvoie (French) (dative-accusative)

Therefore, we can study each clitic cluster as a separate vector space limited in the number of dimensions (sometimes a three-dimensional space can be sufficient). Each space forms a context in which some linguistic phenomena should characterise a variety, that is the vectors of varieties that are similar should be closer in this space. Nevertheless, even similar dialects may have some clitic clusters for which their distance can be high. So how do we measure the distance of two or more dialects when we have many vector spaces? By means of a mathematical operator like combination of subspaces, we can build a higher dimensional space which encloses all the smaller subspaces. An example is the *structured vector space* model which incorporates word meaning in context (Erk and Padó, 2008). On this high dimensional space, we can use the standard definition of cosine similarity among vectors to determine the distance between varieties.

4. References

M. Agosti, P. Benincà, G. M. Di Nunzio, R. Miotto, and D. Pescarini. 2010. A Digital Library Effort to Support the Building of Grammatical Resources for Italian Dialects. In M. Agosti, F. Esposito, and C. Thanos, editors, *Digital Libraries - 6th Italian Research Conference, IRCDL 2010. Revised Selected Papers*, volume 91 of *Communications in Computer and Information Science*, pages 89–100. Springer.

B. Bickel and J. Nichols, 2013. *Obligatory Possessive Inflection*, chapter 58. In Dryer and Haspelmath (Dryer and Haspelmath, 2013).

V. Colonna, A. Boattini, C. Guardiano, I. Dall'Ara, D. Pettener, G. Longobardi, and G. Barbujani. 2010. Long-range comparison between genes and languages based on syntactic distances. *Human Heredity*, 70(4):245–254.

M. S. Dryer and M. Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

K. Erk and S. Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 897–906, Stroudsburg, PA, USA. Association for Computational Linguistics.

R. W. Hamming. 1950. Error detecting and error correcting codes. *Bell System Tech. J.*, 29:147–160.

P. Jaccard. 1901. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547–579.

M. Melucci. 2008. A Basis for Information Retrieval in Context. *ACM Trans. Inf. Syst.*, 26(3):1–14, June.

J. Nerbonne and W. Wiersma. 2006. A Measure of Aggregate Syntactic Distance. In *Proceedings of the Workshop on Linguistic Distances*, LD '06, pages 82–90. Association for Computational Linguistics.

Q. Nguyen, X Nguyen, M. O'Neill, and A. Agapitos. 2012. An Investigation of Fitness Sharing with Semantic and Syntactic Distance Metrics. In A. Moraglio, S. Silva, K. Krawiec, P. Machado, and C. Cotta, editors, *Genetic Programming*, volume 7244 of *Lecture Notes in Computer Science*, pages 109–120. Springer Berlin Heidelberg.

G. Salton, A. Wong, and C. S. Yang. 1975. A Vector Space Model for Automatic Indexing. *Communication of the ACM*, 18(11):613–620, November.

A. Siewierska, 2013. *Passive Constructions*, chapter 107. In Dryer and Haspelmath (Dryer and Haspelmath, 2013).

M. R. Spruit, W. Heeringa, and J. Nerbonne. 2009. Associations among linguistic levels. *Lingua*, 119(11):1624–1642.

M. R. Spruit. 2006. Measuring syntactic variation in dutch dialects. *Literary and Linguistic Computing*, 21(4):493–506.

M. R. Spruit. 2008. *Quantitative perspectives on syntactic variation in Dutch dialects*. LOT Dissertation Series 174. Netherlands Graduate School of Linguistics / Landelijke (LOT).

[2] See examples in "Lectures on Clitics" `http://www.lel.ed.ac.uk/~packema/teaching/ling2L/index.htm`