Biomedical Informatics Insights

ORIGINAL RESEARCH

Decomposing Phenotype Descriptions for the Human Skeletal Phenome

Tudor Groza[1], Jane Hunter[1] and Andreas Zankl[2,3]

[1]School of ITEE, The University of Queensland, Australia. [2]Bone Dysplasia Research Group, UQ Centre for Clinical Research (UQCCR), The University of Queensland, Australia. [3]Genetic Health Queensland, Royal Brisbane and Women's Hospital, Herston, Australia. Corresponding author email: tudor.groza@uq.edu.au

Abstract: Over the course of the last few years there has been a significant amount of research performed on ontology-based formalization of phenotype descriptions. The intrinsic value and knowledge captured within such descriptions can only be expressed by taking advantage of their inner structure that implicitly combines qualities and anatomical entities. We present a meta-model (the Phenotype Fragment Ontology) and a processing pipeline that enable together the automatic decomposition and conceptualization of phenotype descriptions for the human skeletal phenome. We use this approach to showcase the usefulness of the generic concept of phenotype decomposition by performing an experimental study on all skeletal phenotype concepts defined in the Human Phenotype Ontology.

Keywords: human skeletal phenome, phenotype decomposition, phenotype segmentation, ontologies

Biomedical Informatics Insights 2013:6 1–14

doi: 10.4137/BII.S10729

This article is available from http://www.la-press.com.

Introduction

Phenotype descriptions are important for our deeper understanding of genetics and evolutionary relationships. These facilitate the computation and analysis of evolutionary questions related to a varied range of issues, such as the genetic and developmental bases of correlated characters or the paleontological correlates of particular types of change in genes, gene networks, and developmental pathways.[1] The literature contains a wealth of such phenotype descriptions, usually reported as free-text entries. A first and crucial step required to be able to take advantage of this knowledge is to model and capture them in a machine-processable format.

Ontology-based formalization of phenotype descriptions has been a throughly researched topic over the course of the last few years. Consequently, a number of projects have emerged, some of the most representative being the Human Phenotype Ontology (HPO),[2] the Elements of Morphology Project,[3] the Mammalian Phenotype Ontology,[4] or the Phenoscape Project.[5] Applications of formalized phenotypes include the study of cross-species phenotype networks,[6,7] linking human diseases to animal models[8] or predicting diagnoses using semantic similarity measures.[9,10]

However, as noted also by Gkoutos et al,[11] in order to fully capture the intrinsic value and knowledge expressed by these descriptions, we require a more precise and fine-grained representation for them. Most phenotype terms implicitly combine anatomical entities with qualities. For example, HP:0000260 ("wide anterior fontanelle") describes the anatomical entity "anterior fontanelle" that bears the quality "wide." Other terms represent atomic phenotypes that do not externalize this association directly, for example, HP:0010884 ("acromelia"), although their semantics can still be encoded using the same format, that is, use of the explicit meaning of "acromelia" that denotes "shortness of the distal part of a limb." This has led to the emergence of the Entity-Quality (EQ) formalism that enables the decomposition of phenotypic descriptions using ontologies, such as the Foundational Model of Anatomy (FMA),[12] describing anatomical concepts, and the Phenotype and Trait Ontology (PATO),[11] comprising quality definitions. Subsequently, tools for manually creating such associations have been proposed, for example, Phenoscape[5] and Phenex.[13]

In this paper, we advance the state of the art by proposing a holistic solution for the automatic decomposition and conceptualization of phenotype descriptions. This solution relies on two elements: (1) an ontology, the Phenotype Fragment Ontology (PFO), aimed at providing a scaffolding onto which concepts can be created by reusing entities from widely adopted ontologies; and (2) a processing pipeline that takes as input the textual representation of a phenotype description and provides as output its decomposed, ontological representation.

To our knowledge, this represents the first attempt to provide a completely automatic approach for bridging the gap between plain text and logical formalizations of phenotype descriptions. As we will discuss, previous work exists mostly on the manual decomposition of phenotype concepts defined by HPO into logical expressions.[14,15] However, to date, no solution has been proposed to automatically decompose any phenotype description into its elementary units. Furthermore, our processing pipeline can be easily adapted to work with existing formalization solutions, and it is thus not strictly dependent on PFO.

The goal of PFO is to capture the inner structure of phenotype descriptions by enabling the construction of complex phenotypes via combinations of anatomical entities (ie, "epiphysis," part of "phalanx") and qualities (ie, "wide"). Similar to earlier models, for example, Mungall et al,[14,15] PFO provides a metamodel for phenotypes where the actual concepts (ie, anatomical entities and qualities) are defined via well-known and widely adopted ontologies in the biomedical domain, such as FMA and PATO.

The processing pipeline, on the other hand, comprises three steps: (1) segmentation, that is, phenotype descriptions are segmented into the corresponding anatomical entities and qualities; (2) alignment, that is, segmented anatomical entities and qualities are aligned to corresponding concepts in FMA and PATO; and (3) representation, that is, concepts resulting from the alignment phase are used to create PFO entities via class axioms.

Each step of our processing pipeline, and hence of the decomposition process, has associated challenges. Segmentation is affected by ambiguity and lack of a uniform internal textual structure (ie, there is no clear pattern that could be used to denote the anatomical entities and qualities within a phenotype description).

Alignment is influenced by segmentation errors and lexical and terminological differences (eg, "spinal facet joint" vs. "vertebral arch joint"). However, as we will show, the solutions we propose are able to successfully deal with most of these challenges and achieve a high efficiency.

The context of our research is provided by the SKELETOME project,[16] which aims to create a community-driven knowledge curation platform for the skeletal dysplasia domain.[a] To date, we have developed an ontology, the Bone Dysplasia Ontology,[17] capable of capturing associations between skeletal dysplasias, gene mutations, and phenotypic descriptions, the latter grounded in HPO concepts. The decomposition of phenotype descriptions, in our case represented mostly by radiographic findings of the skeletal system, would enable a fine-grained exploration of the phenotype space and, hence, the exploration of commonalities between disorders based on the anatomical localization of phenotypes and the development of anatomical localixation-oriented decision support methods. Consequently, our work focuses on elements associated only with the human skeletal phenome.

To showcase the applicability of the generic concept of decomposing phenotype descriptions, we performed an experimental study on all skeletal phenotypes defined in the Human Phenotype Ontology (3538 terms). The study has enabled an analysis of all elements involved in the decomposition process, starting with our ontological scaffolding and processing pipeline and ending with all ontologies used in the process, that is, HPO, FMA, and PATO. As we will discuss, this analysis has revealed interesting qualitative (ie, terminological issues and missing terms) and quantitative (ie, coverage statistics) findings, in particular, in the context of the existing ontologies.

Methods

The decomposition and conceptualization of phenotype descriptions consists of two main elements: (1) the Phenotype Fragment Ontology, and (2) the processing pipeline. Both are detailed in the following sections.

[a] Bone dysplasias are a group of heterogeneous genetic disorders that affect predominantly the skeletal development. Patients diagnosed with such disorders suffer from complex medical issues that can be described via clinical findings, for example, pains in limbs, radiographic findings, for example, bilateral arachnodactyly and genetic findings, for example, deletion mutation in *FGFR3*.

The Phenotype Fragment Ontology

The Phenotype Fragment Ontology (PFO) aims to provide a standard representation for phenotype descriptions based on their intrinsic structure and to enable the creation of the corresponding entities via class axioms and reused concepts from FMA and PATO. As shown in Figure 1, PFO defines five concepts (depicted with continuous lines) and four relations (depicted with bold lines), all of which are discussed below. In addition, PFO imports one PATO concept (PATO:0000001, "quality"), three FMA concepts (Physical_anatomical_entity, Primary_anatomical_coordinate and Secondary_anatomical_coordinate) and two relations from the Relation Ontology,[18] "has_part" and "part_of."

The central concept of the ontology, Phenotypic_Composite, carries a bridging role between anatomical parts expressed within descriptions and the qualities they bear. Starting from this central concept, the design of PFO introduces elements to accommodate different internal structures that phenotype descriptions may have. The simplest structure associates an existing anatomical concept to an existing quality. This is realized via the "describes" relation, connecting the Phenotypic_Composite to FMA: Physical_anatomical_entity, and the "has_quality" relation that connects the same Phenotypic_Composite to PATO:0000001 ("quality"). These two relations are, in fact, the most important relations introduced by PFO, the others having the role of augmenting composite structures (as discussed below). Considering as example HP:0000260 ("wide anterior fontanelle"), the logical definition in the Manchester syntax would be:

```
Class: Decomposed_HP_0000260
 SubClassOf:
  describes only
  FMA:Anterior_fontanelle
  and describes some
  FMA:Anterior_fontanelle
 SubClassOf:
  has_quality only PATO:0002359
and has_quality some PATO:0002359
 SameAs:
  HP:0000260
```

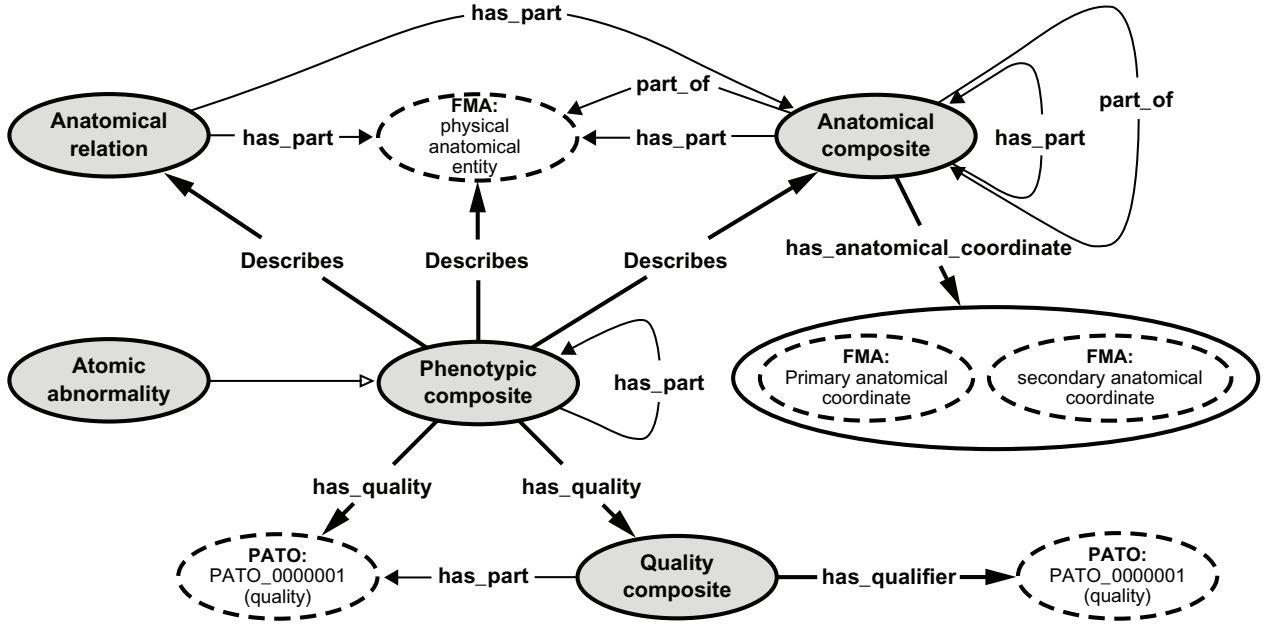

Figure 1. The Phenotype Fragment Ontology.
Notes: Concepts introduced by the ontology are depicted with continuous lines, while those imported from other ontologies, such as FMA and PATO, are depicted with dotted lines. Similarly, the relationships introduced by PFO are bolded in the figure, while those imported from the Relation Ontology are not.

```
 Annotations:
  rdfs:comment "PATO:0002359
(broad): A quality inhering in a
bearer by virtue of the bearer's
width being notably higher than its
length."
```

In order to cater to the more complex structures, the design of PFO introduces additional composite concepts, as listed below:

- Anatomical_Composite, which enables part-subpart relationships between anatomical entities and localization of anatomical coordinates (via the "has_anatomical_coordinate" relationship), for example, epiphysis of proximal, phalanx of 4th toe
- Anatomical_Relation, which allows for anatomical entities to be combined in the context of a quality, that is, "fusion" of hamate and—capitate
- Quality_Composite, which enablis the modelling of composite qualities, including associated qualifiers (via the "has_qualifier" relationship), that is, "delayed" closure of the anterior fontanelle

Below we present the logical definition of the example used also to describe the processing pipeline (see Fig. 2), that is, HP:0100200 ("stippling of the epiphysis of the proximal phalanx of the 4th toe").

```
Class: Decomposed_HP_0100200
 SubClassOf:
  describes only AC_00001
  and describes some AC_00001
 SubClassOf:
  has_quality only PATO:0001512
  and has_quality some PATO:0001512
 SameAs:
  HP:0100200
 Annotations:
  rdfs:label "Stippling of the epi-
physis of the proximal phalanx of
4th toe"
  rdfs:comment "PATO:0001512 (punc-
tate): A pattern inhering in a
surface by virtue of the bearer's
being marked by the presence of
dots or punctures."

Class: AC_00001
 SubClassOf:
  has_part only FMA:Epiphysis and
  has_part some FMA:Epiphysis
```

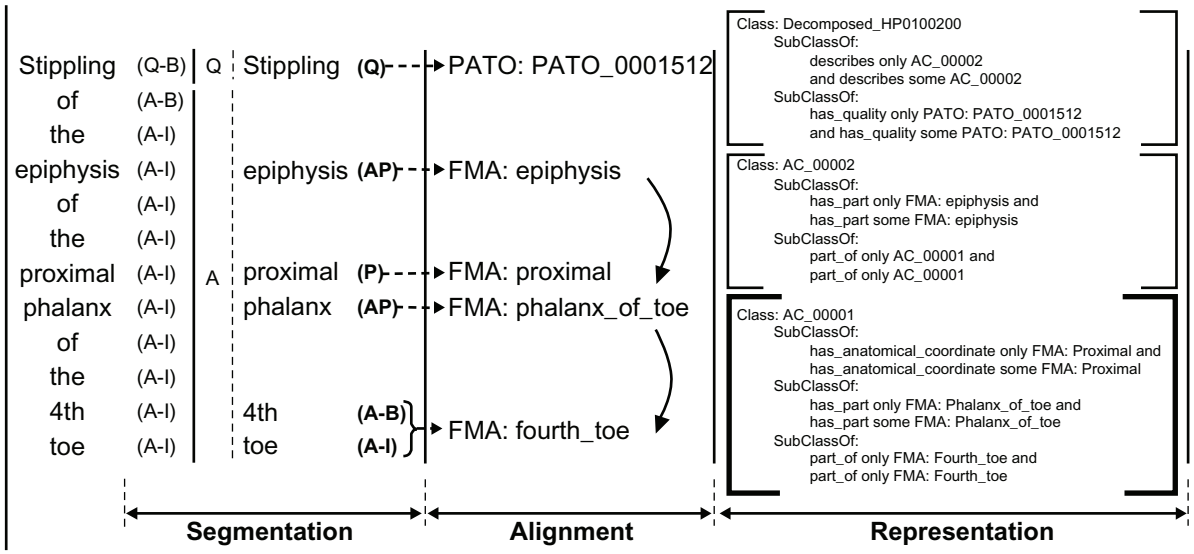

Figure 2. Phenotype decomposition and conceptualization pipeline.
Notes: The pipeline has three phases: 1. Segmentation, that is, textual representations of the phenotype descriptions are segmented into their atomic elements, that is, anatomical (A) and quality (Q) entities, that is, resulting segments are reordered and further segmented into anatomical parts (AP, A) and coordinates (P), and qualities and qualifiers, respectively. Both segmentations use the BIO format; 2. Alignment, that is, resulted segments are aligned to FMA and PATO concepts; part-subpart relationships between anatomical entities are preserved; 3. Representation, that is, aligned concepts are used to create PFO entities.

```
 SubClassOf:
  part_of only AC_00002 and
  part_of only AC_00002

Class: AC_00002
 SubClassOf:
  has_anatomical_coordinate only
FMA:Proximal
  and has_anatomical_coordinate
some FMA:Proximal
 SubClassOf:
  has_part only FMA:Phalanx_of_toe and
  has_part some FMA:Phalanx_of_toe
 SubClassOf:
  part_of only FMA:Fourth_toe and
  part_of only FMA:Fourth_toe
```

As mentioned earlier, not all phenotypes externalize their intrinsic structure. An exception to the general rule is represented by atomic phenotypes. These represent widely adopted terms, usually from Latin, that hide this structure, for example, HP:0010884 (acromelia). PFO considers such terms (modelled via the Atomic_Abnormality concept) to be a subtype of Phenotypic_Composites and, thus, enables their decomposition as in the case of any other phenotype description. Enabling logical definitions for such atomic phenotypes represents one of the major innovations and advantages brought by PFO. To exemplify, below we show the logical definition of the above mentioned concept, "acromelia".

```
Class: Decomposed_HP_0010884
 SubClassOf:
  describes only AC_00001
  and describes some AC_00001
 SubClassOf:
  has_quality only PATO:0000574
  and has_quality some PATO:0000574
 Annotations:
  rdfs:label "Acromelia"
  skos:altLabel "Short distal part
of the limb"
  rdfs:comment "PATO:0000574
(decreased length): A length qual-
ity which is relatively small."
 SameAs:
  HP:0010884

Class: AC_00001
 SubClassOf:
  has_part only FMA:Limb and
  has_part some FMA:Limb
```

```
 SubClassOf:
  has_anatomical_coordinate only
FMA:Distal
  and has_anatomical_coordinate
only FMA:Distal
```

In addition to the conceptual definitions, in order to provide a proper ontological context for entities defined by PFO, all classes are rooted in entities defined by the Basic Formal Ontology (BFO, http://www.ifomis.org/bfo),[19] and by the Ontology of General Medical Science (OGMS),[20] a middle ontology rooted in BFO, which provides a specific framework for medicine. The concept mappings are listed in the following: (1) Phenotypic_Composite, Atomic_Anomaly and Quality_Composite represent OGMS:0000023 ("phenotype"), which is a snap:Quality, and (2) Anatomical_Composite and Anatomical_Relation are snap:MaterialEntity, because they define material anatomical entities.

As mentioned, previous work exists on creating logical expressions for phenotype descriptions (see Mungall et al[14,15]), and a direct mapping can be drawn between some aspects of PFO and formalizations previously proposed (eg, the "has_qualifier" relation is present in both, or the "towards" relation can be expressed in terms of part-subpart relationships with Anatomical_Composites). There are, however, two major differences: (1) in PFO, we've opted for named concepts, rather than anonymous instances, to enable an easier reuse and analysis of composite elements, and (2) we've provided a clear mechanism for identifying anatomical coordinates, as well as defining atomic phenotypes, which may have a vague anatomical localization. Nevertheless, the processing pipeline has been designed in a flexible manner, so that other formalization approaches could be adopted.

Automatic decomposition and conceptualization pipeline

Our decomposition and conceptualisation pipeline consists of three phases: (1) segmentation, in which textual representations of the phenotype descriptions are segmented into their atomic elements, anatomical and quality entities; (2) alignment, in which resulting segments are aligned to FMA and PATO concepts; and (3) representation, in which concepts resulting from alignment are used to create PFO entities.

Figure 2 depicts the pipeline by means of an example, HP:0100200 ("stippling of the epiphysis of the proximal phalanx of the 4th toe"). The remainder of this section uses the same example to discuss each phase of the pipeline.

Segmentation

The segmentation of phenotype descriptions raises a series of structural and semantic challenges. From a structural perspective, there are four classes of segments that need to be considered: qualities, qualifiers, anatomical coordinates, and anatomical entities, the latter being decomposable into parts and sub-parts. Considering the example in Figure 2, "stippling" denotes the quality, while "phalanx" denotes an anatomical entity and has associated an anatomical coordinate ("proximal"). Secondly, due to the composite nature of the anatomical concepts, there is no uniform pattern that can be assumed for segmentation. For example, "epiphyseal widening of the hand phalanges" is the same as "broadening of the epiphyses of the phalanges of the hand." From a semantic perspective, one challenge is provided by ambiguity, for example, "irregular ossification of the proximal radial metaphisis" versus "radial club hand," where "radial" refers to the anatomical entity, "radius," in the first case and to an anatomical coordinate in the second case. Finally, the existing terminology contains metaphorical expressions that may pose issues for an accurate detection/classification, for example, "bone-in-bone appearance" or "angel-shaped epiphyses."

Machine learning methods have proved to be successful at dealing with the above mentioned challenges,[21] although rule-based methods could also be employed with a high precision, but most likely at a trade-off of a lower recall.[22] Conditional Random Fields (CRF),[23] in particular, have been reported to achieve good results both for segmentation tasks as well as for classification tasks in the biomedical domain.[24,25] Recent works, however, have concentrated on using ensembles of classifiers (hybrid approaches) to overcome the issues associated with using single classifiers. As an example, Zhou et al[26] and Torii et al[27] have used sets of classifiers (three by the former and six by the latter) aggregated via different voting schemes for gene/protein mention tagging.

Our solution also relies on training divergent models via an ensemble of classifiers. Additionally, in

order to improve the segmentation results, we have experimented with multiple aggregation schemes, such as set operations and simple majority voting. Each aggregation technique has been the subject of an individual experiment. Overall, we have adopted a two-phase process, as exemplified in Figure 2. Firstly, we segment the input into coarse qualities and anatomical entities using the BIO format. Then we reorder these coarse-grained segments according to their class and split them into atomic parts. These atomic elements correspond to quality-qualifier pairs (eg, "delayed—closure"), anatomical coordinate-anatomical concept associations and part-sub part relationships between anatomical concepts (eg, epiphysis—of—phalanxof—4th toe—see Fig. 2). In both phases, our ensemble of classifiers comprises two CRFs and two Support Vector Machines (SVM)-based chunkers. The features used for classification are listed in Table 1.

Alignment

The result of the segmentation phase is a set of text chunks with associated labels (as exemplified in Fig. 2), that is, Q = quality, QF = qualifier, AP = anatomical part, P = anatomical coordinate, and A = main anatomical entity. The goal of this phase is, for each segment, to find the best corresponding candidate in FMA and PATO, respectively (subject to its type). In practice, this reverts to a lexical similarity task where any of the existing similarity metrics could be employed.

In order to increase the efficiency of this alignment phase, instead of performing a direct similarity comparison between segments and labels/synonyms of ontological concepts, we created similarity matrices

Table 1. Features used for classification in the segmentation phase. Examples are provided using the token "epiphysis" from Figure 2.

Feature	Description	Example
Token	Current token	epiphysis
Token prefix (variable size)	Token prefixes (size in example is 5)	e ep epi epip epiph
Token postfix (variable size)	Token postfixes (size in example is 5)	s is sis ysis physis
Token shape	Shape of token by replacing all capitalized letters with 'A', all non- capitalized letters with 'a' and all digits with '0'	aaaaaaaaa
Token brief shape	Compressed version of the token shape where all consecutive equal characters are compressed	a
Token lemma	Token lemma (stem)	epiphysi
Token POS tag	Part of speech tag of token	NNP
Morpho: punctuation	Flag to indicate whether the token ends in a punctuation sign	no
Morpho: vowels	Shape of token provided by replacing all consonants with '-'	e-i----i-
Morpho digits	Shape of token by replacing all digits with '*'	no*
Context: unigram	Unigram-based surrounding context of token (variable window size). Window size in example is 3	Stippling of the epiphysis of the proximal
Context: bigram	Bigram-based surrounding context of token (variable window size). Window size in example is 3	Stippling-of of-the the-epiphysis epiphysis-of of-the the-proximal
Context: trigram	Trigram-based surrounding context of token (variable window size). Window size in example is 3	Stippling-of-the of-the-epiphysis the-epiphysis-of epiphysis-of-the of-the-proximal
Dictionary: conjunctions	Lexicon comprising conjuctions (and, or)	
Dictionary: connectives	Lexicon comprising connective tokens (at, of, the, etc)	
Dictionary: ordinals	Lexicon comprising ordinals (1st, 2nd, etc)	
Dictionary: coordinates	Lexicon comprising anatomical coordinates (central, left, etc)	
Dictionary: anatomy	Gazzetteer compiled from unigrams of FMA concepts	
Dictionary: quality	Gazzetteer compiled from unigrams of PATO concepts	

and adapted the concept of matrix trace (which can be computed only for squared matrices and represents the sum of the diagonal elements) to fit our needs. Figure 3 depicts a concrete example using the segment "vertebral bodies" and two FMA concepts, Spinal_reticular_process and Body_of_vertebra.

The exact alignment steps are the following:

- We pre-process both the input segment and the label/synonyms of the concept candidate by transforming them into lower case, removing punctuation, performing tokenization and removing tokens that represent stop words (see Part B of Fig. 3).
- For each pair segment-concept label/synonym, we create two similarity matrices, one using the tokens of the segment in the normal order and one by reversing their order. Part A of Figure 3 depicts the similarity matrix created using the normal order of *vertebral bodies* and FMA:Spinal_reticular_process, while part B of the same figure depicts the similarity matrix created using the reverse order of "vertebral bodies" and FMA:Body_of_vertebra. Each cell of the similarity matrix is computed using equation 1, where NLCS is normalized longest common subsequence (LCS) and NMCLCS is the normalized maximal consecutive longest common subsequence starting at 1 (ie, with the first character) and respectively at n (ie, starting anywhere in the string).
- Within each similarity matrix we compute traces for all square submatrices by performing an arithmetic mean over their diagonals. The final trace of the similarity matrix is the maximum subtrace computed for all submatrices with diagonal N, where N is the length of the given segment in tokens (N = 2 in our example).
- The final similarity is computed via Eq. 5, using the maximum between the trace of the similarity matrix of the normal order and the trace of the similarity matrix of the reverse order (ie, MaxTrace). The first component of Eq. 5 represents a penalty factor influenced by the difference in length between the given segment and the ontological concept, with N = number of tokens in the segment and L = number of tokens in the lexical representation of the ontological concept. In our example, N = 2 and L is 3 in part A and 2 in part B of Figure 3. The goal of this component is to penalize the ontological concepts that use only some of its tokens in the computation of the similarity matrix.

$$\begin{aligned} sim(s_1, s_2) = {} & w_1 * NLCS(s_1, s_2) \\ & + w_2 * NMCLCS_1(s_1, s_2) \\ & + w_3 * NMCLCS_n(s_1, s_2) \end{aligned} \quad (1)$$

$$NLCS(s_1, s_2) = \frac{length(LCS(s_1, s_2))^2}{length(s_1) * length(s_2)} \quad (2)$$

$$NMCLCS_1(s_1, s_2) = \frac{length(MCLCS_1(s_1, s_2))^2}{length(s_1) * length(s_2)} \quad (3)$$

$$NMCLCS_n(s_1, s_2) = \frac{length(MCLCS_n(s_1, s_2))^2}{length(s_1) * length(s_2)} \quad (4)$$

$$Sim = \max \left\| \left(e^{-\frac{|N-L|}{N}} - \frac{|N-L|}{N * e} \right) * MaxTrace \right\| \quad (5)$$

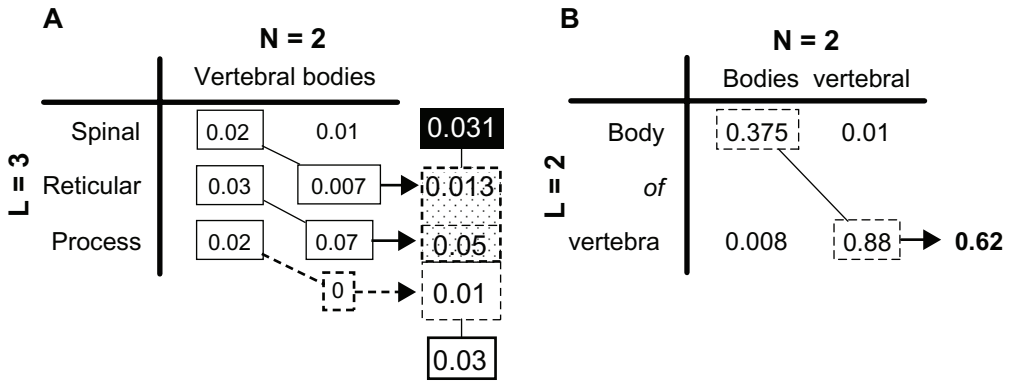

Figure 3. Similarity matrix and traces computation. (**A**) Similarity matrix and traces computation between vertebral bodies using the normal order and the FMA concept Spinal_reticular_process; (**B**) Similarity matrix and traces computation between the inverse order of vertebral bodies and the FMA concept Body_of_vertebra; stop-words are discarded during the traces computation.

Once the best concept candidate is selected for each segment of the phenotype description (ie, the candidate with the maximum similarity score), we perform one last step, preserving the part-subpart relationship between anatomical entities. As depicted in Figure 2, under the Alignment phase, our aim is to link those anatomical concepts that form a part-subpart relationship, for example, Epiphysis—of—Phalanx_of_toe—of—Fourth_toe. This step is realized by using the FMA structure, and more concretely, the "rdfs:subClassOf and fma:constitutional_part" relations.

In practice, starting from the main anatomical entity of the phenotype description (labeled A as a result of the second phase of segmentation), we employ a modified version of Dijkstra's algorithm for shortest paths in a graph (using simultaneously both types of edges, that is, subclass and constitutional_part) to find the paths between this and all the other segmented anatomical parts. The anatomical part with the shortest path is then marked as the new main anatomical entity and the process is repeated until all anatomical parts are linked. In our example, we start from Fourth_toe and find that the shortest path is to Phalanx_of_toe. Since there's only one anatomical part left, that is, Epiphysis, the algorithm infers the part-subpart relationship listed above. However, if more than one anatomical parts would have been available, the process would have marked Phalanx_of_toe as the new main anatomical entity and would have repeated the previous step.

Representation

This last phase of our processing pipeline reassembles the segmented and aligned phenotype description into a logical definition. We have defined a set of rules that enable us to create PFO entities using the results of the alignment phase, that is, FMA and PATO concepts, and the part-subpart relationships, where relevant. Considering the example depicted in Figure 2, we follow a bottom-up approach and start by creating the Anatomical_Composite (ie, AC_00001) and the class axioms that describes the relationship between Proximal, Phalanx_of_toe and Fourth_toe, then we create AC_00002 following the same rules, and finally create the main Phenotypic_Composite that connects logically AC_00002 to the quality PATO:0001512. It is easy to observe that in order to adopt a different formalization scheme (eg, Mungall et al[14,15]), it is enough to create a set of rules corresponding to that formalization with the rest of the pipeline remaining unchanged.

Results

In order to understand the capabilities of both the ontology as well as of the processing pipeline, we've performed an experiment on all skeletal phenotypes defined by the Human Phenotype Ontology (v17.07.2012). These are represented by 3538 concepts, that is, the subtree of HP:0000924 ("abnormality of the skeletal system") and account for more than a third of the entire ontology. The experiment has followed the steps associated with the processing pipeline and was finalized with a qualitative analysis of the resulting concepts. In this section, we detail the results achieved by the processing pipeline, or more specifically, by the first two steps, while in the following section, we discuss our findings. As a remark, the same experiment can also be performed on concepts defined by the Mammalian Phenotype Ontology or on phenotypes defined by RADLEX (http://www.radlex.org/). Our focus on the HPO skeletal phenotypes has been strictly motivated by the needs emerging from the SKELETOME project, as already mentioned in Introduction.

Segmentation Results

As discussed in the previous section, the segmentation phase consists of two steps (see also Fig. 2): (1) the initial segmentation of the phenotype descriptions into coarse anatomical and quality elements; and (2) the fine-grained segmentation of these coarse elements into quality–qualifier concepts and anatomical parts–anatomical coordinates entities. Both steps have been performed using an ensemble of classifiers that comprises two CRFs and two Support Vector Machines (SVM)-based chunkers. The specific packages used to train the classifiers were: (1) CRF++ (http://crfpp.googlecode.com/), a freely available CRF package that we have used to train a forward parsing model; (2) MALLET,[28] another freely available CRF package that we have used to train a forward parsing model; and (3) YamCha,[29] a chunking package that uses SVM classification and we have trained two models that differ in the method used for

the multiclass classification, that is, one versus one or one versus all.

The actual experiments have been carried out on all 3538 HPO concepts via a 5-fold cross-validation with stratification. The HPO concepts have been manually segmented and tagged to form the corpus used for cross-validation. We used different aggregation strategies, ranging from simple set operations (that is, union and intersection of results), composite set operations (aggregated pairs of set operations, that is, the union of two classifiers intersected with the union of the other two classifiers) to a simple majority voting scheme with veto (since we had an even number of classifiers). In order to measure the classification efficiency, we have used the standard precision, recall, and F-1 score. The results reported below represent the average scores achieved within the cross-validation process.

Table 2 summarizes the best results for step 1 of the segmentation of the best individual classifier, as well as each of the aggregation techniques. In addition to noting the high efficiency achieved by all strategies (97.04% F-1 with dictionaries and 96.77% without dictionaries), there are two particular aspects that are worth discussing. Firstly, the difference in efficiency between using or omitting the dictionaries is minimal. This proves that our classifiers do not require any external sources in order to deliver accurate results. Secondly, within each class, we can observe a minimal difference also between the aggregation strategies. For example, YamCha1vsAll, as an individual classifier, has performed on par with the best set operation scheme and only slightly worse (0.10% difference) than the voting mechanism (with YamCha1vs1 as veto holder).

Table 3 lists the best results for the second step of the segmentation phase. Each class has been evaluated individually; however, the general comparative remark across aggregation strategies remains valid. In the anatomy category, we can observe that an individual classifier (in this case CRF++ and YamCha1vsAll, 97.15% F-1) has again performed better than the best set operation strategy (96.74% F-1) and slightly worse than the voting mechanism, 97.26% F-1, with CRF++ as veto holder. Exactly the same trend is visible also in the quality category. Cross-category we can see that the classifiers have performed worse in the quality category, which was an expected behavior, in principle due to the atomic abnormalities. (See also the discussion in the following section.)

Overall, our experiments lead to the conclusion that using an ensemble of classifiers for segmentation tasks may not necessarily improve the overall accuracy because of its dependency on the goal and underlying data characteristics. Hence, in this particular case, opting for a single classifier without dictionaries provides the best trade-off between accuracy and the amount of features used for classification plue the complexity of the classification architecture.

Alignment Results

The alignment phase was evaluated against all anatomical and quality concepts that had a correspondence in FMA and PATO. In total, the experiment has been carried out on 330 anatomical concepts and 183 quality concepts represented in phenotype segments by 538 anatomical tokens and 328 quality tokens, respectively (ie, some concepts had more than one lexical representation due to synonyms or terminological differences, while others had a single standard representation).

Table 4 lists the evaluation results. The alignment F-1 scores were 87.17% for anatomical concepts and 91.56% for quality concepts–without using external sources (eg, dictionaries) or human feedback. A closer look at the alignment output has revealed the following issues: (1) in most cases, the anatomical alignment has failed due to

Table 2. Comparative results for the first step of the segmentation phase.

Method	With dictionaries			Without dictionaries		
	P (%)	R (%)	F-1 (%)	P (%)	R (%)	F-1 (%)
YamCha1vsAll	96.94	96.94	96.94	96.70	96.70	96.70
Set operations	96.63	97.21	96.92	96.27	97.03	96.65
Voting (YamCha1vs1)	97.04	97.04	97.04	96.77	96.77	96.77

Table 3. Comparative results for the second step of the segmentation phase.

Method	Anatomy			Quality		
	P (%)	R (%)	F-1 (%)	P (%)	R (%)	F-1 (%)
CRF++/YamCha1vsAll	97.15	97.15	97.15	92.21	92.21	92.21
Set operations	96.74	96.74	96.74	91.32	91.32	91.32
Voting (CRF++/CRF++)	97.26	97.26	97.26	92.44	92.44	92.44

terminological differences. For example, HPO concepts use terms, such as, iliac wings or vertebral facet, while FMA models them as "ala of ilia" and "verterbral arch." Hence, without human intervention or precomputed dictionaries, such alignments cannot be automatically performed.

There were, however, also cases where our alignment strategy has failed, for example, where HPO terms omit parts of the actual anatomical concept, unilamboid/bilambdoid (referring to a single or both lambdoid sutures, however, used without the explicit presence of the token "suture") versus "lambdoid suture," where the chosen candidate has been Lambda; (2) the vast majority of improper quality alignments have been on "metaphoric" concepts, that is, concepts consisting of shape comparisons and ending in "-shaped", for example, "Y-shaped" or "pear-shaped." Table 7 shows that these are commonly used qualities. A second issue has been found in aligning terms that end in "-ing" (eg, "cupping") to their corresponding concepts ending in "-ed" (eg, "cupped"). In practice, we've observed a general tendency of HPO using terms that refer to "processes" (eg, cupping, ossification, maturation) rather than their associated qualitative results (eg, cupped, ossified, matured).

Discussion

The experimental study performed on the HPO skeletal phenotypes has provided us with insights that go beyond understanding the efficiency achieved by our method. Firstly, it reveals the distribution of FMA and PATO concepts, as well as the coverage of the missing FMA and PATO concepts in these skeletal phenotype descriptions. As it can be seen in Table 5, almost 14% of anatomical terms described by the phenotypes are not present in FMA, although in some instances these can be found in RADLEX. Examples include "acetabular roof," "cerebellar dendate nucleus," "mandibular rami" (RADLEX:RID28576), "sacroiliac notch," "sacrosciatic notch," "vertebral endplates" (RADLEX:RID6126), "Talar dome" (RADLEX:RID2954), and, in general, most elements related to "dermatoglyphs." On the quality side, the same table shows that there is a fair balance between atomic phenotypes (ie, phenotypes that do not externalize their internal structure) and missing and existing qualities in PATO. The large majority of missing qualities are "metaphoric" descriptions, such as "swan neck-like" or "chevron-shaped." From a coverage perspective (see Table 6), missing FMA concepts are present in around 8% of the HPO concepts, while missing qualities and atomic phenotypes account for around 36% of the HPO concepts (19% atomic phenotypes and 17% missing qualities).

Secondly, it allows us to build a view over the commonly occurring concepts, as presented in Table 7. We can observe that 30% of the HPO skeletal phenotypes describe abnormalities of the "phalanges" (30%), followed by "epiphysis" (17%), "toe" (17%) and "finger" (15%). As a remark, these concepts are not mutually exclusive since there may be HPO

Table 4. Evaluation results of the alignment phase.

	Precision (%)	Recall (%)	F-1 (%)
Anatomy	88.81	85.59	87.17
Qualities	93.05	90.13	91.56

Table 5. Distribution of FMA and PATO concepts in the skeletal phenotypes in HPO.

	Anatomy (FMA)	Quality (PATO)
Existing concepts	330 (86.16%)	183 (33.27%)
Non-existing concepts	53 (13.83%)	165 (30%)
Atomic phenotypes	–	202 (36.72%)

Table 6. Coverage of missing FMA and PATO concepts in the skeletal phenotypes in HPO.

Concepts	Coverage
Non-existing FMA concepts	275 (7.77%)
Atomic phenotypes	683 (19.30%)
Non-existing PATO concepts	592 (16.73%)

concepts describing composite anatomical entities (eg, "epiphysis of the phalanx of toe"). The table also shows the most often occurring anatomical coordinates, that is, "proximal (11%) and "distal" (10%). On the qualities side, the concept "abnormality" is the most common (in 9% of the terms), followed by "hypoplasia," "aplasia," and metaphoric descriptions ending in "-shaped," all with around 5%.

Finally, it enables an analysis of the qualitative differences between terminologies and a deeper understanding of the shortfalls of our alignment strategy. The alignment of HPO terms to FMA concepts is not trivial due to several terminological and structural differences. FMA follows an uniform terminological structure, usually subpart-part, for example, Body_of_vertebra, while HPO is closer to the terminology used in clinical practice and the medical literature (eg, "vertebral bodies" and "calf muscles"). In practice, these issues could be addressed by enriching concepts with appropriate synonyms and/or adopting an uniform scientific terminology (ie, "superficial muscle of posterior compartment of "leg" instead of "calf muscles"). Here, we could also point out that HPO uses nonspecific (or context dependent) terms that do not necessarily have a direct anatomical correspondence, which makes the alignment problematic. The most representative example is the term "phalanx" (or "phalanges") that require a proper context (ie, "finger" or "toe") in order to be aligned to the corresponding FMA concept (ie, Phalanx_of_finger or Phalanx_of_toe). A final set of terminological differences are related to the mixtures of languages. Both HPO and FMA comprise English and Latin terms; however, FMA uses the Latin representation usually as a synonym and provides an English counterpart where this exists, while HPO sometimes omits this English counterpart, even if the literature provides one, for example, pectus carinatum or "pigeon breast."

Table 7. Often occurring concepts in the skeletal phenotypes in HPO.

Anatomy	Quality
Phalanx (1094—30.92%)	Abnormality (337—9.52%)
Epiphysis (612—17.29%)	Hypoplasia (187—5.28%)
Toe (598—16.9%)	Aplasia (178—5.03%)
Finger (537—15.17%)	-shaped (174—4.91%)
Proximal (384—10.85%)	Sclerosis (136—3.84%)
Distal (367—10.37%)	Duplication (133—3.75%)
Middle (278—7.85%)	Absent (102—2.88%)

From our perspective, we see the following major open challenges: (1) defining and representing atomic phenotypes, a complex task that could be achieved by analyzing the textual description that accompanies them; (2) defining and representing abnormalities that involve relationships between anatomical elements, abnormalities of specific parts of anatomical elements (eg, fingertips or interdigital folds), and abnormalities of spatial, functional and nonfunctional properties of anatomical elements (eg, mineral density, movement, angles); and (3) improving the alignment of anatomical concepts by using external resources, such as the textual definition of the phenotype descriptions.

Conclusion

In this paper, we have proposed a complete solution for decomposing phenotype descriptions in an automatic manner. Our approach consists of the following two elements: (1) an ontology, the Phenotype Fragment Ontology, that enables the definition of complex phenotypes, via class axioms, by reusing concepts from FMA and PATO; and (2) a processing pipeline that "segments" phenotype descriptions into their atomic entities (ie, anatomic and quality elements), "aligns" the resulted entities to FMA and PATO concepts by preserving the existing part-subpart relationships between the anatomical entities, and creates the corresponding ontological "representation" according to PFO.

The experimental results have showed that each step of the processing pipeline achieves a high accuracy. Furthermore, the analysis enabled by the decomposition of the HPO skeletal phenotypes led to a series of interesting findings, ranging from missing concepts in FMA and PATO to terminological difference between HPO and FMA and shortfalls of our Phenotype Fragment Ontology.

Future work will focus on improving PFO and the processing pipeline and using our approach to align skeletal phenotypes in all existing phenotype ontologies.

Funding

This research has been funded by the Australian Research Council (ARC) under the Discovery Early Career Researcher Award (DECRA)-DE120100508 and the Linkage grant SKELETOME-LP100100156.

Competing Interests

Author(s) disclose no potential conflicts of interest.

Author Contributions

Conceived and designed the experiments: TG. Analyzed the data: TG and JH. Wrote the first draft of the manuscript: TG. Contributed to the writing of the manuscript: JH and AZ. Agree with manuscript results and conclusions: TG, JH, and AZ. Jointly developed the structure and arguments for the paper: TG. Made critical revisions and approved final version: TG. All authors reviewed and approved of the final manuscript.

Disclosures and Ethics

As a requirement of publication author(s) have provided to the publisher signed confirmation of compliance with legal and ethical obligations including but not limited to the following: authorship and contributorship, conflicts of interest, privacy and confidentiality, and (where applicable) protection of human and animal research subjects. The authors have read and confirmed their agreement with the ICMJE authorship and conflict of interest criteria. The authors have also confirmed that this article is unique and not under consideration or published in any other publication, and that they have permission from rights holders to reproduce any copyrighted material. Any disclosures are made in this section. The external blind peer reviewers report no conflicts of interest.

References

1. Mabee PM, et al. Phenotype ontologies: the bridge between genomics and evolution. *Trends Ecol Evol*. 2007;22(7):345–50.
2. Robinson PN, et al. The Human Phenotype Ontology: A tool for annotating and analyzing human hereditary disease. *Am J Hum Genet*. 2008;83(5): 610–5.
3. Carey JC, et al. Standard terminology for phenotypic variations: The elements of morphology project, its current progress, and future directions. *Hum Mutat*. 2012. [Epub ahead of print.]
4. Smith C, Goldsmith C, Eppig J. The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol*. 2005;6(1):R7.
5. Dahdul WM, et al. Evolutionary characters, phenotypes and ontologies: Curating data from the systematic biology literature. *PLoS One*. 2010;5(5): e10708.
6. Hoehndorf R, Schofield PN, Gkoutos GV. PhenomeNET: A whole-phenome approach to disease gene discovery. *Nucleic Acids Res*. 2011;39(18):e119.
7. Schofield PN, Hoehndorf R, Gkoutos GV. Mouse genetic and phenotypic resources for human genetics. *Hum Mutat*. 2012;33(5):826–36.
8. Washington NL, et al. Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biol*. 2009;7(11):e1000247.
9. Kohler S, et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am J Hum Genet*. 2009;85(4):457–64.
10. Paul R, Groza T, Zankl A, Hunter J. Semantic similarity-driven decision support in the skeletal dysplasia domain. Presented at: *The 11th International Semantic Web Conference*; Nov 11–5, 2012; Boston, MA.
11. Gkoutos GV, et al. Entity/Quality-based logical definitions for the human skeletal phenome using PATO. *Proceedings of 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*; Sep 3–6, 2009; Minneapolis, MN. 7069–72.
12. Rosse C, Mejino JLV. A reference ontology for biomedical informatics: The Foundational Model of Anatomy. *J Biomed Inform.* 2003;36(6): 478–500.
13. Balhoff JP, et al. Phenex: Ontological annotation of phenotypic diversity. *PLoS One*. 2010;5(5):e10500.
14. Mungall C, Gkoutos G, Washington N, Lewis S. Representing phenotypes in OWL. In: *Proceedings of the OWLED 2007 Workshop on OWL: Experiences and Directions*; Jun 6–7, 2007; Innsbruck, Austria.
15. Mungall C, Gkoutos G, Smith CL, Haendel MA, Lewis SE, Ashburner M. Integrating phenotype ontologies across multiple species. *Genome Biol*. 2010;11(1):R2.
16. Groza T, Zankl A, Yuan-Fang L, Hunter J. Using Semantic Web technologies to build a community-driven knowledge curation platform for the skeletal dysplasia domain. In: *Proceedings of the 10th International Semantic Web Conference*; Oct 23–7, 2011; Bonn, Germany. 81–96.
17. Groza T, Hunter J, Zankl A. The Bone Dysplasia Ontology: integrating genotype and phenotype information in the skeletal dysplasia domain. *BMC Bioinformatics*. 2012;13:50.
18. Smith B, Ceusters W, Klagges B, et al. Relations in biomedical ontologies. *Genome Biol*. 2005;6(5):R46.
19. Grenon P, Smith B, Goldberg L. Biodynamic Ontology: Applying BFO in the biomedical domain. *Stud Health Technol Inform*. 2004;102:20–38.
20. Scheuermann RH, Ceusters W, Smith B. Toward an ontological treatment of disease and diagnosis. In: *Proceedings of the AMIA Summit on Translational Bioinformatics*; Mar 15–7, 2009; San Francisco, CA. 116–20.
21. Cesario E, Folino F, Locane A, Manco G, Ortale R. Boosting text segmentation via progressive classification. *Knowl Inf Syst*. 2008;15(3):285–320.
22. Cortez E, da Silva AS, Golcalves MA, Mesquita F, de Moura ES. FLUX-CIM: Flexible unsupervised extraction of citation metadata. In: *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*. New York, NY: ACM; 2007:215–24.
23. Lafferty JD, McCallum A, Pereira FCN. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In: *Proceedings of the Eighteenth International Conference on Machine Learning*. San Francisco, CA: Morgan Kaufmann Publishers Inc; 2001:282–9.
24. Sun C, Guan Y, Wang X, Lin L. Rich features based Conditional Random Fields for biological named entities recognition. *Comput Biol Med*. 2007;37(9):1327–33.
25. Li L, Zhou R, Huang D. Two-phase biomedical named entity recognition using CRFs. *Comput Biol Chem*. 2009;33(4):334–8.
26. Zhou G, Shen D, Zhang J, Su J, Tan S. Recognition of protein/gene names from text using an ensemble of classifiers. *BMC Bioinformatics*. 2005;6(Suppl 1):S7.

27. Torii M, Hu Z, Wu C, Liu H. BioTagger-GM: a gene/protein name recognition system. *J Am Med Inform Assoc*. 2009;16(2):247–55.
28. McCallum AK. A Machine Learning for Language Toolkit. MALLET. 2002. http://mallet.cs.umass.edu. Accessed December 21, 2012.
29. Kudoh T, Matsumoto Y. Use of support vector learning for chunk identification. In: *CoNLL '00 Proceedings of the 2nd workshop on learning language in logic and the 4th conference on computational natural language learning, volume 7*. Stroudsburg, PH; Association for Computational Liguistics; 2000:142–4.