MultiMASC: An Open Linguistic Infrastructure for Language Research

Nancy Ide

Department of Computer Science
Vassar College, USA
ide@cs.vassar.edu

Abstract

This paper describes MultiMASC, which builds upon the Manually Annotated Sub-Corpus (MASC) (Ide et al., 2008; Ide et al., 2010) project, a community-based collaborative effort to create, annotate, and validate linguistic data and annotations on a broad-genre open language data. MultiMASC will extend MASC to include comparable corpora in other languages that not only represent the same genres and styles, but also include similar types and number of annotations represented in a common format. Like MASC, MultiMASC will contain only completely open data, and will rely on a collaborative community-based effort for its development. We describe the possible ways in which additional corpora for MultiMASC can be collected and annotated and consider the dimensions along which "comparability" for MultiMASC corpora can be determined. Because it is unlikely that all language-specific MultiMASC corpora can be comparable along every dimension, we also outline the measures that can be used to gauge comparability for a number of different criteria.

Keywords: Comparable corpora, Corpus construction, Multi-lingual resources

1. Introduction

In an ideal universe, computational linguistics researchers would have open access to very large language corpora spanning the full range of genres, registers, and languages, all of which would be accompanied by high quality annotations for linguistic phenomena at all levels that can be used to support machine learning and computational linguistics research in general. Parallel data would exist for all languages, and common lexical, semantic, and discourse-level phenomena would be linked across data of all genres and languages. Annotations would come with detailed information about provenance as well as evaluation metrics in order to ensure quality, and researchers could easily request specific data and annotations to be delivered as needed over the web, in a physical format and using "annotation semantics" that can be integrated without modification into their own tools and resources. Unfortunately, this scenario is a long way off, and the greatest obstacle is the high cost of high-quality resource production and maintenance. Another obstacle is the difficulty of obtaining language data representing a variety of genres that is unfettered by licensing constraints so that it may be used for any purpose community-wide.

The Manually Annotated Sub-Corpus (MASC) (Ide et al., 2008; Ide et al., 2010) project attempts to overcome these obstacles to high-quality resource creation through a community-based collaborative effort to create, annotate, and validate linguistic data and annotations on broad-genre open language data. MASC is a half million word corpus of contemporary American English language data drawn from the 15 million word Open American National Corpus (OANC)[1] that includes manually produced or validated annotations for a wide range of linguistic phenomena at all linguistic levels. The corpus includes a balanced set of nineteen genres of spoken and written language data that is completely open for any use. The corpus is freely downloable from the MASC website, as well as through the Linguistic Data Consortium (LDC)[2]. All MASC annotations are represented in a common format so that they may be used collectively to study intra-level interactions, which are important for the deeper analyses that are increasingly the focus in the field.

This paper describes MultiMASC, which builds upon the MASC project by extending MASC to include comparable corpora in other languages. Here, "comparable" means not only representing the same genres and styles, but also include similar types and number of annotations represented in a common format. Like MASC, MultiMASC will contain only completely open data and rely on a collaborative community-based effort for its development.

We first describe MASC as it currently exists, as well as plans for its future development. The remainder of the paper describes the possible ways in which additional corpora for MultiMASC can be collected and annotated. We then consider the dimensions along which "comparability" for MultiMASC corpora can be determined, and, because it is unlikely that all language-specific MultiMASC corpora can be comparable along every dimension, we outline the measures that can be used to gauge comparability for a number of different criteria.

2. MASC

MASC is the only corpus with multiple layers of annotations in a common format that can be used either individually or together, and (unlike, for example, OntoNotes) to which others can add annotations. MASC will be soon increased in size to a million words, although there are currently no resources for further in-house validation; we will depend on the community to validate and contribute annotations to fill in the gap.

[1] http://www.anc.org/OANC

[2] http://www.ldc.upenn.edu

MASC currently contains nineteen genres of spoken and written language data in roughly equal amounts, shown in Table 1. Approximately 15% of the corpus consists of spoken transcripts, both formal (court and debate transcripts) and informal (face-to-face, telephone conversation, etc.); the remaining 85% covers a wide range of written genres, including emerging social media genres (tweets, blogs). Because it is drawn from the OANC, all MASC data represents contemporary American English produced since 1990. The entire MASC is annotated for logical structure, token and sentence boundaries, part of speech and lemma, shallow parse (noun and verb chunks), named entities (person, location, organization, date), and Penn Treebank syntax. Portions of MASC are also annotated for additional phenomena, including 40K of full-text FrameNet frame element annotations and PropBank, TimeML, and opinion annotations over a roughly 50K subset of the data. As the name of the corpus implies, all annotations have either been manually produced or automatically produced and hand-validated. The list of annotation types and coverage is given in Table 2.

MASC also includes sense-tags for 1000 occurrences of each of 100 words chosen by the WordNet and FrameNet teams (100,000 annotated occurrences), described in (Passonneau et al., 2012). The sense-tagged data are distributed as a separate *sentence corpus* with links to the original documents in which they appear. Where MASC does not contain 1000 occurrences of a given word, additional sentences were drawn from the OANC. Several inter-annotator agreement studies and resulting statistics have been published (Passonneau et al., 2009; Passonneau et al., 2010), many of which are distributed with the corpus.

Genre	No. files	No. words	Pct corpus
Court transcript	2	30052	6%
Debate transcript	2	32325	6%
Email	78	27642	6%
Essay	7	25590	5%
Fiction	5	31518	6%
Gov't documents	5	24578	5%
Journal	10	25635	5%
Letters	40	23325	5%
Newspaper	41	23545	5%
Non-fiction	4	25182	5%
Spoken	11	25783	5%
Technical	8	27895	6%
Travel guides	7	26708	5%
Twitter	2	24180	5%
Blog	21	28199	6%
Ficlets	5	26299	5%
Movie script	2	28240	6%
Spam	110	23490	5%
Jokes	16	26582	5%
TOTAL	376	506768	

Table 1: Genre distribution in MASC

All MASC annotations are represented in the ISO TC37 SC4 Linguistic Annotation Framework (LAF) GrAF format (Ide and Suderman, 2007; Ide and Suderman, Submitted), with the objective to make the annotations as flexible for use with common tools and frameworks as possible. The ANC project provides a web application, called ANC2Go[3] that enables a user to choose any portion or all of MASC and the OANC together with any of their annotations to create a "customized corpus" that can be delivered in any of several widely used formats such as CONLL IOB, RDF, inline XML, etc. Modules to transduce GrAF to formats consistent with other tools and frameworks such as UIMA, GATE, and NLTK are also provided.[4] Thus "openness" in MASC applies to not only acquisition and use, but also interoperability with diverse software and systems for searching, processing, and enhancing the corpus.

Annotation type	No. words
Logical	506659
Token	506659
Sentence	506659
POS/lemma (GATE)	506659
POS (Penn)	506659
Noun chunks	506659
Verb chunks	506659
Named Entities	506659
FrameNet	39160
Penn Treebank	*506659
PropBank	55599
Opinion	51243
TimeBank	*55599
Committed Belief	4614
Event	4614
Dependency treebank	5434

* under development

Table 2: Summary of MASC annotations

3. MultiMASC

MultiMASC will both expand MASC and the collaboration effort upon which it depends and exploit the infrastructure and experience that the development of MASC has provided. The eventual result will be a massive, multilingual, multi-genre corpus with comparable multilayered annotations that are inter-linked via reference to the original MASC, as shown in Figure 1.

We see the development of MultiMASC as an incremental process, involving the following steps for any given language:

1. Create and make available a corpus of open language data, comparable in size and genre distribution to MASC.
2. Collect and make available annotations for linguistic phenomena comparable to, and possibly extending beyond, those available for MASC, either automatically or manually produced, in any format.
3. Validate the automatically-produced annotations.
4. Provide the annotations in a format compatible with MASC and other MultiMASC annotations.

[3] http://www.anc.org:8080/ANC2Go/

[4] http://www.anc.org/tools/

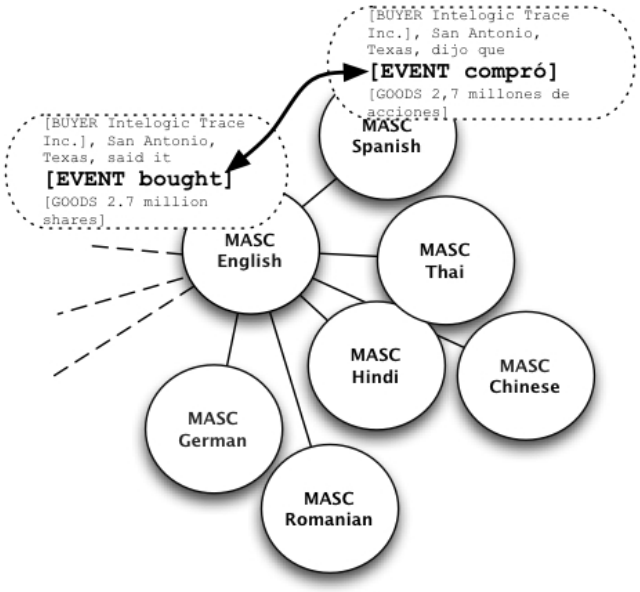

Figure 1: Overview of MultiMASC

5. Provide linkage among annotations in the language-specific data and MASC annotations, and/or annotations in other MultiMASC corpora as appropriate.

Given the expected constraints of funding and resources, we anticipate that for some languages, interim results will be all that is available at any given point in development, or, possibly, that interim results are all that ever becomes available. Even if this is the case, the comparable MultiMASC corpora created in step 1 will provide a resource for computational linguistics research and development that is unmatched at present.

4. Step one: Data gathering

The first step in the creation of MultiMASC is to produce a massive multi-lingual corpus of language-specific data with comparable genre distribution that is open and freely available for community use. "Open" in OANC/MASC terms means that data is either in the public domain or under a license that does not restrict redistribution of the data or its use for any purpose, including commercial use (e.g., the Creative Commons Attribution license[5]). Data under licenses such as GNU General Public License[6] or Creative Commons Attribution-ShareAlike[7] are avoided because of the potential obstacles to use for commercial purposes imposed by the requirement to redistribute under the same terms.

Comparable MultiMASC component corpora will need approximately 25,000 words of open data for each of the nineteen MASC genres, produced by native speakers of the language in question (no translations) after 1989. Fortunately, experience shows us that obtaining and preparing samples of this size is considerably easier than for larger amounts of data, which will hopefully make the prospect of constructing a language-specific portion of MultiMASC less daunting for potential contributors.

4.1. Obtaining open data

The OANC/MASC project has long been identifying and gathering open data for inclusion in both the OANC and MASC. The following are some of the sources and strategies we have utilized:

1. Contributions from publishers who are willing to provide data under a non-restrictive license, as is the case for the OANC/MASC non-fiction materials donated by Oxford University Press and Cambridge University Press, and SLATE magazine articles from Microsoft. To protect their interests, the publishers sometimes provided only a subset of a complete book or collection.

2. Web search for materials in the public domain. Government documents and debate and court transcripts, as well as technical articles in collections such as Biomed and PLOS, are typically in the public domain, for example.

3. Web search for data licensed under non-viral licenses such as CC-BY. Blogs, fiction, and other writing such as essays are very often distributed over the web under these terms.

4. Contributions from college students of class essays and other writing. College students produce considerable volumes of prose during their academic careers, and very often this data is discarded or forgotten once handed in to satisfy an assignment. The OANC site provides a web interface for contributions of this kind

[5] http://creativecommons.org/licenses/by/3.0/
[6] http://www.gnu.org/licenses/gpl.html
[7] http://creativecommons.org/licenses/by-sa/2.5/

that includes a grant of permission to use the contributed materials. We regularly solicit these contributions from students in our own and other institutions.

5. Contributions of data from colleagues in the field. We have received data contributions, including significant amounts of spoken data, from several NLP and linguistics projects. As awareness of the need for open data increases, such contributions should become easier to obtain.

6. Direct solicitation for use of web materials. We have on occasion identified a web site containing interesting or substantial materials and contacted the relevant parties directly to explain our use of the data and ask for permission to use it. We have also contacted providers whose data are freely available for access to the materials in a form more manageable for processing purposes. So far, none of our requests has been turned down.

Different languages, as well as different countries and therefore different copyright laws, will affect the ease with which MultiMASC data can be acquired in any given case. To the extent that it applies, the experience of the MASC project can be relied upon as a resource to support the acquisition of MultiMASC data.

4.2. Identifying comparable data

The definition of "comparable" as it applies to genre is, of course, not exact. The best guideline to determine comparability may be to consider the primary uses to which MultiMASC will be put, including the extraction and/or linkage of parallel segments and paraphrases; semantic frame elements; translations of single words, multi-word expressions, proper names, and named entities; etc., in order to facilitate inter-linguistic discoveries and comparisons. To address this, we can identify several dimensions along which to measure cross-lingual comparability, including structural complexity; lexical richness and specificity; vocabulary register; temporal organization (tense and aspect); referential cohesion; interactiveness; and others (see, for example, the measures outlined in (Biber, 1995)).

Statistics characterizing these dimensions (e.g., simple measures such as type/token ratio, word and sentence length, together with metrics indicating the degree of use of linguistic features such as private verbs, suasive verbs, time and place adverbials, subordination, third person pronouns, proper nouns, and many more), which are available for MASC data, may provide a point of departure for determining comparability. However, more research into this possibility will be required to determine exactly what the best among such measures may be, and, more critically, how the measures may or may not apply depending on the language in question.

Beyond comparability on the basis of metrics like these, we may also consider comparability in terms of topic, that is, data that treats the same or a closely related topic as the original MASC document. One possibility is to consider a continuum of comparability, starting with the most general: same domain (e.g. finance), same topic (e.g., investment), same sub-topic (e.g., 401K accounts), same subject (e.g., report or description of same event, etc.).

4.3. Preparing the data

The ANC project has extensive experience in preparing data that is obtained in any of several formats for use by annotation tools. This experience can be exploited by developers of MultiMASC component corpora in order to make the data preparation process easier, if not entirely trivial. For example, we have an automatic pipeline for processing documents originally in Microsoft Word, Open Office (odt), or Rich Text Format (rtf) that generates a UTF-8 file containing the text content together with standoff annotations for logical structure down to the level of paragraph. The annotations can be automatically rendered in any of several possible output formats, including GrAF.

The ANC project has also developed several modules for the General Architecture for Text Engineering (GATE)[8] to import from and export to GrAF, so that annotations generated within GATE can be immediately rendered in the MASC common format. GATE includes annotation modules for a fairly extensive range of languages, which means that in some cases, generating automatically-produced annotations for MultiMASC in GrAF will be trivial. We have also developed similar GrAF import/export modules for the UIMA annotation framework.

5. Step two: Annotation

Getting the MultiMASC data in place for as many languages as possible provides the base for a community effort to annotate the data. For major languages, it should be relatively easy to obtain automatically-produced annotations comparable to the basic MASC annotations: sentence and token boundaries, at least one part-of-speech/lemma analysis, shallow parse (noun and verb chunks), syntactic phrase structure (trees), and basic named entities (person, organization, location, date).

Validation of the annotations is a much more costly and time-intensive venture. MASC validation has so far been done in-house by trained validators; however, this may not always be feasible, and it is therefore expected that for MultiMASC, considerably more community-based collaboration may be required. The range of possibilities include, at one end, simply publishing the data and unvalidated annotations for community use, with the request that those who use the data contribute any correction or additional annotation they perform.[9] At the other extreme, a sophisticated web-based interface could be provided so that others can directly validate the data, which would track and evaluate annotations as they are produced, use active learning to suggest possible corrections, etc. Crowdsourcing, with or without a sophisticated interface, provides another alternative.

Beyond the types of annotation included (e.g., part-of-speech, named entities, etc.), annotations will ideally be comparable in terms of *syntactic interoperability*, i.e.,the physical format in which they are represented e.g., inline

[8] http://gate.ac.uk

[9] This is the strategy used for the OANC.

vs.standoff annotation, XML, Penn Treebank-like bracketing, etc.). To ensure that all annotations on all language data are usable together and/or with the same tools, annotations can be rendered in the common format used by MASC (LAF/GrAF), or in a format that is trivially mapped to GrAF.

Semantic interoperability among annotations, which involves the actual categories and features used to describe the various linguistic phenomena, is far more difficult to achieve. Clearly, the use of common annotation categories among MultiMASC corpora is not feasible, given that most annotations will first be produced using existing software, and re-tooling existing software to accommodate specific annotation categories (even if it were possible to specify a definitive set that would accommodate all languages and linguistic theories) is unrealistic. Efforts such as ISOCat[10], which attempt to provide ways to map semantic categories and, where this is not possible, specify their differences, are underway. This may enable a greater degree of semantic interoperability among MultiMASC corpora, but such efforts are not expected to be well enough along in the next few years to provide a comprehensive solution. The best measure of comparability that may be possible in the near term might be an indication of the "mappability" between two schemes on a rough scale of difficulty (trivial, medium, hard, unmappable). Ideally, where possible, mappings between schemes for like annotation types among languages would be developed and distributed from the MultiMASC home website.

6. Step three: Creating the inter-linked MultiMASC

The final step in creating MultiMASC will be to link like annotations across languages. We envision linkage among linguistic phenomena at many levels, e.g., part-of-speech categories, syntactic structures, paraphrases, semantic roles, named entities, events, etc.

Linkage among the MultiMASC corpora can be accomplished in at least two ways. First, MASC can be used as a "hub", as depicted in Figure 1, to which annotations of the same phenomenon (a "buy" event in the figure) are directly linked.[11] We anticipate that MultiMASC corpora will be represented in GrAF or a format that is trivially mappable to GrAF. Inter-linkage is then straightforward: an attribute can be added to the XML element for an annotation in a MultiMASC corpus that refers to a corresponding annotation in the American English MASC.

A more elegant and workable solution for inter-linkage among MultiMASC corpora would utilize a reference set of categories, possibly represented in RDF/OWL (for example, resources included in the Linguistic Linked Open Data cloud[12]) and/or residing in a data category registry such as ISOCat[13]. In this scenario, annotations in both MASC and other MultiMASC corpora are linked to an independent entity on the web that provides information about the annotation content, as depicted in Figure 2. For example, a "noun plural" part-of-speech annotation in MultiMASC corpora could include a reference to the PID (persistent identifier) in the ISOCat registry that defines this category. In GrAF, such a reference could look like this:[14]

```
<a label="Token" ref="ann-n3" as="xces">
 <fs>
  <f name="msd" value="...DC-3581"/>
  ...
```

Linkage of this nature will enable cross-linguistic and inter-layer studies on a scale that is currently impossible. Available multi-lingual data from sources such as Wikipedia does not include the layers of annotation we envision for MultiMASC, and Wikipedia data is not completely open due to the restriction to "share-alike". The recently launched Language Library effort[15] includes multiple annotations, but it includes only a handful of materials, most also under "share-alike" constraints, and there is no effort to provide annotations in compatible formats or to inter-link them.

7. Comparability Index

We seek to identify measures of comparability along the several dimensions outlined above that can be used both as a guidelines for the construction of MultiMASC corpora in other languages and as a gauge of comparability for these corpora once they become a part of MultiMASC. The latter is important because we cannot expect that it will be possible in all or even most cases to conform to a strict set of comparability guidelines; with these measures, users will have information that can inform cross-lingual studies that use the MultiMASC data.

Table 3 shows the various dimensions of comparability and an overview of the measures that will be defined to classify them. Note that in principle, all measures apply to the entire language-specific corpus except for DOMAIN/TOPIC/SUBJECT, which will in most cases apply to individual documents or groups of documents within a specific genre. We can envision ultimately providing a very large matrix giving pair-wise comparability indexes for all languages in MultiMASC.

8. Conclusion

A community-wide, collaborative effort to produce high quality annotated corpora is one of the very few possible ways to address the high costs of resource production, and to ensure that the entire community, including large teams as well as individual researchers, has access and means to use these resources in their work. The OANC and MASC already lay the groundwork for such an effort for English, and extending it to other languages seems a logical next step.

[10] http://www.isocat.org

[11] Note that the use of MASC as a hub does not preclude linkage among other language pairs.

[12] http://linguistics.okfn.org/resources/llod/

[13] http://www.isocat.org

[14] Due to space limitations the ISOcat URI prefix http://www.isocat.org/datcat has been replaced by ellipses.

[15] http://www.languagelibrary.eu

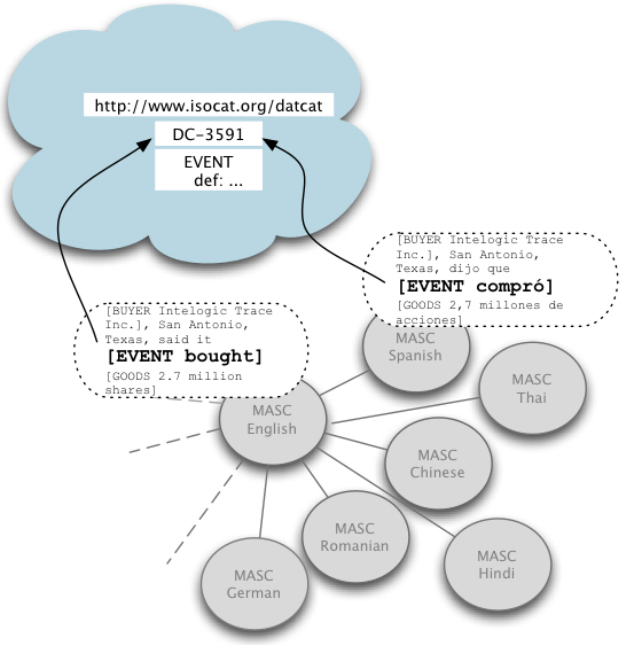

Figure 2: Linked annotations in MultiMASC

Dimension	Information
GENRE	Category among MASC genres
	Comparison measures for each genre, including broad dimensions such as structural complexity, lexical richness and specificity, vocabulary register, etc., with relevant statistics for specific measures (type/token ratio, subordination, use of specific verb types, etc.)
~~DOMAIN/TOPIC/SUBJECT~~*	Continuum along comparability of domain, topic, (one or more) sub-topics, subject
ANNOTATIONS	Comparison with original MASC annotations in terms of the annotation types included, categories provided for each annotation type
	Comparison with annotations included in other language corpora in MultiMASC
	Format, in terms of mappability to a common format or format directly usable with other language corpora in MultiMASC
	Semantics, in terms of conformance or mappability to those in other language corpora in MultiMASC
INTER-LINKAGE	Number and type of inter-linked phenomena

* Applies to individual documents

Table 3: Comparability measures for MultiMASC

The vision of a MultiMASC for a large number and wide variety of languages is to some extent "pie-in-the-sky", as it is certain to take many years to accomplish. Therefore, in order to keep the project within realistic bounds, the plan is to develop MultiMASC opportunistically, incorporating language-specific corpora as they become available and adding annotations and linkages later, if necessary. This way, the community can use and enhance data and annotations as they become available in an extended effort that will hopefully build momentum as the possibilities MultiMASC offers for research become increasingly apparent.

Acknowledgments

This work was supported by National Science Foundation grants CRI-0708952 and CRI-1059312.

9. References

Douglas Biber. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press.

Nancy Ide and Keith Suderman. 2007. GrAF: A graph-based format for linguistic annotations. In *Proceedings of the First Linguistic Annotation Workshop*, pages 1–8, Prague.

Nancy Ide and Keith Suderman. Submitted. The Linguistic Annotation Framework: A Standard for Annotation In-

terchange and Merging. *Language Resources and Evaluation*.

Nancy Ide, Collin Baker, Christiane Fellbaum, Charles Fillmore, and Rebecca Passonneau. 2008. MASC: The Manually Annotated Sub-Corpus of American English. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, Paris. European Language Resources Association.

Nancy Ide, Collin Baker, Christiane Fellbaum, and Rebecca Passonneau. 2010. The Manually Annotated Sub-Corpus : A community resource for and by the people. In *Proceedings of the The 48th Annual Meeting of the Association for Computational Linguistics*, Morristown, NJ, USA, July. Association for Computational Linguistics.

Rebecca J. Passonneau, Ansaf Salleb-Aouissi, and Nancy Ide. 2009. Making sense of word sense variation. In *SEW '09: Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 2–9, Morristown, NJ, USA. Association for Computational Linguistics.

Rebecca Passonneau, Ansaf Salleb-Aouissi, Vikas Bhardwaj, and Nancy Ide. 2010. Word sense annotation of polysemous words by multiple annotators. In *Proceedings of the Seventh Language Resources and Evaluation Conference (LREC)*, Paris. European Language Resources Association.

Rebecca J. Passonneau, Collin Baker, Christiane Fellbaum, and Nancy Ide. 2012. The masc word sense sentence corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, Paris. European Language Resources Association.